



National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file - Votre référence

Our file - Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

UNIVERSITY OF ALBERTA

CAPILLARY GEL ELECTROPHORESIS FOR DNA SEQUENCING

BY

HEATHER R. STARKE

**A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of
the requirements for the degree of Doctor of Philosophy.**

DEPARTMENT OF CHEMISTRY

**Edmonton, Alberta
FALL 1994**



National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file - Votre référence

Our file - Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-95270-9

Canada

Name HEATHER STARKE

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Analytical Chemistry
SUBJECT TERM

0486
SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION

General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0488
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523

Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language
General 0679
Ancient 0289
Linguistics 0290
Modern 0291
Literature
General 0401
Classical 0294
Comparative 0295
Medieval 0297
Modern 0298
African 0316
American 0591
Asian 0305
Canadian (English) 0352
Canadian (French) 0355
English 0593
Germanic 0311
Latin American 0312
Middle Eastern 0315
Romance 0313
Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy 0422
Religion
General 0318
Biblical Studies 0321
Clergy 0319
History of 0320
Philosophy of 0322
Theology 0469

SOCIAL SCIENCES

American Studies 0323
Anthropology
Archaeology 0324
Cultural 0326
Physical 0327
Business Administration
General 0319
Accounting 0272
Banking 0770
Management 0454
Marketing 0338
Canadian Studies 0385
Economics
General 0501
Agricultural 0503
Commerce-Business 0505
Finance 0508
History 0509
Labor 0510
Theory 0511
Folklore 0358
Geography 0366
Gerontology 0351
History
General 0578

Ancient 0579
Medieval 0581
Modern 0582
Black 0328
African 0331
Asia, Australia and Oceania 0332
Canadian 0334
European 0335
Latin American 0336
Middle Eastern 0333
United States 0337
History of Science 0585
Law 0398
Political Science
General 0615
International Law and Relations 0616
Public Administration 0617
Recreation 0814
Social Work 0452
Sociology
General 0626
Criminology and Penology 0627
Demography 0938
Ethnic and Racial Studies 0631
Individual and Family Studies 0628
Industrial and Labor Relations 0629
Public and Social Welfare 0630
Social Structure and Development 0700
Theory and Methods 0344
Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture
General 0473
Agronomy 0285
Animal Culture and Nutrition 0475
Animal Pathology 0476
Food Science and Technology 0359
Forestry and Wildlife 0478
Plant Culture 0479
Plant Pathology 0480
Plant Physiology 0817
Range Management 0777
Wood Technology 0746
Biology
General 0306
Anatomy 0287
Biostatistics 0308
Botany 0309
Cell 0379
Ecology 0329
Entomology 0353
Genetics 0369
Limnology 0793
Microbiology 0410
Molecular 0307
Neuroscience 0317
Oceanography 0416
Physiology 0433
Radiation 0821
Veterinary Science 0778
Zoology 0472
Biophysics
General 0786
Medical 0760
EARTH SCIENCES
Biogeochemistry 0425
Geochemistry 0996

Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences 0768
Health Sciences
General 0556
Audiology 0300
Chemotherapy 0992
Dentistry 0567
Education 0350
Hospital Management 0769
Human Development 0758
Immunology 0982
Medicine and Surgery 0564
Mental Health 0347
Nursing 0569
Nutrition 0570
Obstetrics and Gynecology 0380
Occupational Health and Therapy 0354
Ophthalmology 0381
Pathology 0571
Pharmacology 0419
Pharmacy 0572
Physical Therapy 0382
Public Health 0573
Radiology 0574
Recreation 0575

Speech Pathology 0460
Toxicology 0383
Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences
Chemistry
General 0485
Agricultural 0749
Analytical 0486
Biochemistry 0487
Inorganic 0488
Nuclear 0738
Organic 0490
Pharmaceutical 0491
Physical 0494
Polymer 0495
Radiation 0754
Mathematics 0405
Physics
General 0605
Acoustics 0986
Astronomy and Astrophysics 0606
Atmospheric Science 0608
Atomic 0748
Electronics and Electricity 0607
Elementary Particles and High Energy 0798
Fluid and Plasma 0759
Molecular 0609
Nuclear 0610
Optics 0752
Radiation 0756
Solid State 0611
Statistics 0463
Applied Sciences
Applied Mechanics 0346
Computer Science 0984

Engineering
General 0537
Aerospace 0538
Agricultural 0539
Automotive 0540
Biomedical 0541
Chemical 0542
Civil 0543
Electronics and Electrical 0544
Heat and Thermodynamics 0348
Hydraulic 0545
Industrial 0546
Marine 0547
Materials Science 0794
Mechanical 0548
Metallurgy 0743
Mining 0551
Nuclear 0552
Packaging 0549
Petroleum 0765
Sanitary and Municipal 0554
System Science 0790
Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY

General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

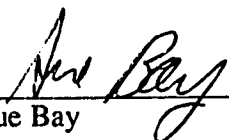



The undersigned grant permission that the material published in

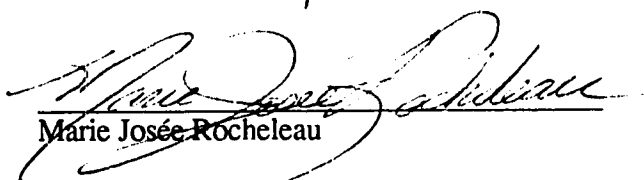
Journal of Chromatography, 1992, 608, 143-150

entitled: "Effect of total percent polyacrylamide in capillary gel electrophoresis for DNA sequencing of short fragments: A phenomenological model"

be used in the Ph.D. thesis of Heather R. Starke


Sue Bay


Jian Zhong Zhang


Marie Josée Rocheleau

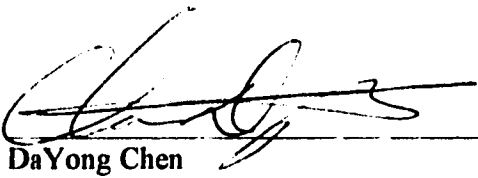

Norman J. Dovichi

The undersigned grant permission that the material published in

Nucleic Acids Research, 1992, 20 4873-4880

entitled: "Two-label peak-height encoded DNA sequencing by capillary gel electrophoresis: three examples"

be used in the Ph.D. thesis of Heather R. Stone


DaYong Chen


Norman J. Dovichi

The undersigned grant permission that the material published in

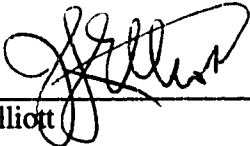
Advances in DNA Sequencing, SPIE, Vol. 1891 (1993), p8.

entitled: "Accuracy of two-color peak height encoded DNA sequencing by capillary gel electrophoresis and laser-induced fluorescence"


be used in the Ph.D. thesis of Heather R. Starke



Sue Bay



John Elliott



Norman Dovichi

The undersigned grant permission that the material submitted for publication in

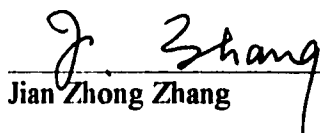
Nucleic Acids Research

entitled: "Internal fluorescence labeling with fluorescent deoxynucleotides in two-label peak-height encoded DNA sequencing by capillary electrophoresis"

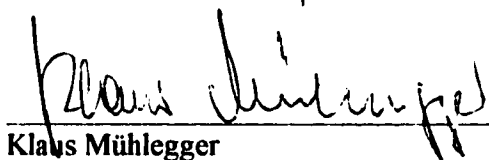
be used in the Ph.D. thesis of Heather R. Starke



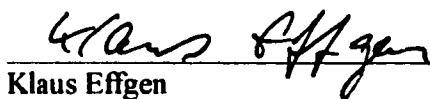
Ju Ying Yan



Jian Zhong Zhang



Klaus Mühlegger



Klaus Effgen



Norman J. Dovichi

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Heather R. Starke
TITLE OF THESIS: Capillary Gel Electrophoresis for DNA Sequencing
DEGREE: Doctor of Philosophy
YEAR THIS DEGREE GRANTED: 1994

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material from whatever without the author's prior written permission .

Heather Starke

9830 - 72 Ave
Edmonton, Alberta, T6E 0Z1

Date: 6 October 1994

UNIVERSITY OF ALBERTA
FACULTY OF GRADUATE STUDIES AND RESEARCH

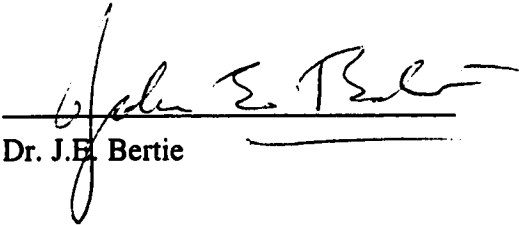
The undersigned certify that they have read, and recommended to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Capillary Gel Electrophoresis for DNA Sequencing** submitted by **Heather R. Starke** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.



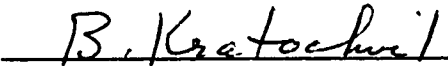
Dr. N.J. Dovichi



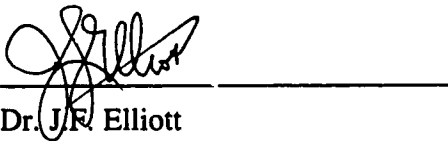
Dr. F.F. Cantwell



Dr. J.E. Bertie



Dr. B.G. Kratochvil



Dr. J.F. Elliott



Dr. B. Koop

Abstract

The advent of large scale DNA sequencing projects drives the continued development of sequencing technologies. Capillary gel electrophoresis (CGE), coupled with a highly sensitive laser induced fluorescence detector, is a high speed technique that is used to separate fluorescently tagged DNA sequencing samples. The behaviour of short fragments (less than 250 bases) is described as a function of total percent acrylamide (%T) in gels with constant crosslinker concentration (5%C). At each concentration peak spacing was constant. The peak spacing increased linearly with the total percent acrylamide. Theoretical plate counts are limited by longitudinal diffusion, and are independent of total percent acrylamide.

A two-label peak-height encoded sequencing technique that is a variation on the peak-height method of Tabor and Richardson was demonstrated. The samples generated with the ABI FAM and JOE labelled primers were readily detected on the CGE instrument constructed for use with the DuPont Genesis sequencing system. A peak height ratio of 2:1 gave sequencing accuracies of 98%, while a peak height ratio of 3:1 gave accuracies of more than 99%.

The performance of the CGE using the two-colour peak-height ratio method was evaluated. Sequence data was generated for five samples of malaria DNA on the CGE instrument and the Pharmacia ALF sequencer. The sequence data was compared to the consensus sequence. The length of the sequence data obtained per run ranged from 300-460, and the accuracy of the sequences obtained by CGE ranged from 97.3-99.8%. The lengths and accuracies were comparable to those from the automated sequencer. The CGE instrument produces the data about 3 times faster than the Pharmacia instrument.

Internal labelling of the sequencing samples was accomplished with fluorescently tagged dUTP and dATP. Fluorescein and tetramethylrhodamine labelled dATPs were used to generate two-colour peak-height encoded samples. The detector was modified to detect these two labels: two lasers, a green HeNe and an Argon ion laser were used to excite the tetramethylrhodamine and fluorescein respectively. Internal labelling provides an simple, inexpensive alternative to primer labelling.

Acknowledgements

Capillary electrophoresis experiments are not carried out in a vacuum. Many people contributed in many ways to the work described in this thesis.

First and foremost, I would like to express my sincere appreciation to Dr. Norman "hands of a sturgeon" Dovichi. From first to last he has been a real encouragement and help. I have learned a great deal from him during the last five years and his gift for choosing research topics is reflected in this thesis.

I would also like to mention all of the past and present members of the Northern Lights Laser Lab. Most of them contributed in some way to this project specifically, and to my overall education. We have shared a lot of experiences and bad puns over the years, we have experienced growing pains as the group expanded, and we have become good friends as well as colleagues. Specifically, I owe a great deal to Dr. Harold Swerdlow, who started the DNA research and guided us through the early work. Dr. David Chen bailed me out of countless problems in the lab. I have always appreciated his quiet patience and encouragement. Sue Bay, summer student extraordinaire, was responsible for a great deal of the work presented in chapters 2 and 4. It was a joy to work with Sue.

The excellent support staff in the machine shop, the electronics shop, and the offices also deserve mention. Without them to keep things running smoothly this work would have taken much longer.

My sincere thanks to Dr. John Elliott and Dean Smith of the Departments of Medical Microbiology and Infections Diseases and Immunology, for hands on training in the fine art of sequencing. One slab gel was enough to convince me that capillaries held a lot of promise for the field. Dr. Satyender Hansra provided us with the malaria clones as well as her sequence data for the comparison in Chapter 4.

Dr. Carl Fuller of United States Biochemical Corporation also contributed immensely to this work, providing reagents and enzymes, technical expertise and many entertaining telephone conversations.

Finally, I would like to thank my parents, Siegfried and Margaret Harke, and my husband Dieter Starke for their unfailing love and encouragement.

Table of Contents

CHAPTER 1 Introduction	1
1.1 The Structure of DNA.....	1
1.2 DNA Sequencing Techniques.....	5
1.2.1 Chain Extension and Termination - Sanger's Method.....	6
1.2.2 The Maxam-Gilbert Chemical Degradation Technique.....	8
1.3 Modified Sequencing Approaches.....	13
1.3.1 One Fluorescent Primer.....	13
1.3.2 Four Fluorescent Primers.....	14
1.3.3 Dye Terminators.....	14
1.3.4 Internal Labelling with Fluorescent dNTPs.....	15
1.3.5 Peak Height Methods.....	16
1.3.6 Cycle Sequencing.....	16
1.4 Mapping and Sequencing the Genome.....	18
1.4.1 Restriction Digests.....	18
1.4.2 Gel Electrophoresis.....	20
1.4.3 Cloning.....	21
1.4.4 Mapping and Sequencing.....	21
1.5 Capillary Electrophoresis.....	23
1.6 Laser-Induced Fluorescence	26
References	32
CHAPTER 2 The Effect of Total Percent Polyacrylamide in Capillary Gel Electrophoresis for DNA Sequencing of Short Fragments.....	35
2.1 Introduction	35
2.2 Experimental.....	36

2.2 Experimental.....	36
2.2.1 Instrument Design.....	36
2.2.2 Gel Preparation.....	40
2.2.3 Sample Preparation.....	41
2.2.4 Electrophoresis.....	41
2.3 Results and Discussion.....	42
2.3.1 Performance and Stability of the Gel Filled Capillaries.....	42
2.3.2 Sequencing Rate and Retention Time.....	47
2.3.3 Electrophoretic Mobility.....	51
2.3.4 Theoretical Plates.....	55
2.3.5 Band Broadening - injection, detection, and thermal gradient.....	55
2.3.6 Band Broadening - longitudinal diffusion.....	58
2.3.7 Resolution.....	60
2.4 Conclusions.....	63
References.....	65
 CHAPTER 3 Two-Label Peak Height Encoded DNA Sequencing by Capillary	
Gel Electrophoresis.....	67
3.1 Introduction.....	67
3.2 Experimental.....	69
3.2.1 Electrophoresis.....	69
3.2.2 Detector.....	69
3.2.3 Sample Preparation.....	71
3.2.4 Sequence Determination.....	72
3.3 Results and Discussion.....	72
3.4 Conclusions.....	94

References	96
CHAPTER 4 Accuracy of Two-Colour Peak Height Encoded DNA	
Sequencing by Capillary Gel Electrophoresis and Laser-Induced Fluorescence	97
4.1 Introduction	97
4.2 Experimental	99
4.2.1 Instrument Design.....	99
4.2.2 Gel Preparation.....	99
4.2.3 Sample Preparation.....	100
4.2.4 Electrophoresis.....	101
4.2.5 ALF Sequencing.....	102
4.3 Results and Discussion	102
4.3.1 Sequence Accuracy.....	103
4.3.2 Sequence Length	119
4.3.3 Sequencing Rate	120
4.3.4 Resolution	120
4.4 Conclusions.....	121
References	123
CHAPTER 5 Internal Fluorescence Labelling with Two-Label Peak-Height	
Encoded DNA Sequencing by Capillary Gel Electrophoresis.....	124
5.1 Introduction	124
5.2 Experimental.....	127
5.2.1 Instrument Design.....	127
5.2.2 Detector	128
5.2.3 Sample Preparation.....	128
5.2.4 Electrophoresis.....	132
5.3 Results and Discussion	132

5.3.1 F-dUTP Labelled Sample.....	132
5.3.2 Two-Colour dATP Labelled Sample	133
5.3.3 The Labelling Reaction	137
5.3.4 Comparison of Separation of Internal and Primer Labelled Samples.....	146
5.4 Conclusions.....	155
References	156
CHAPTER 6 Conclusions and Future Work	157
6.1 Conclusions.....	157
6.2 Future Work	158
References	160
Appendix A Consensus Sequences.....	161
1. M13mp18	161
2. Mouse mammary tumour virus - MMTV.....	164
3. Mouse cytokine - 123 delta	165
4. Malaria templates 1 to 5	166
Appendix B Problems in DNA Sequencing.....	171
1. Compressions.....	171
2. Non-specific Priming.....	173
3. Degraded Template DNA.....	176

List of Tables

Table 1.1 Chemical Modifications for Maxam-Gilbert Sequencing.....	10
Table 1.2 Some restriction enzymes and their recognition sites.....	19
Table 4.1 Length and accuracy of DNA sequence determined by CGE and automated slab gel electrophoresis.....	118

List of Figures

Figure 1.1	2
a. A Nucleotide: the basic repeating unit of DNA	2
b. The purine and pyrimidine bases	2
Figure 1.2 Single stranded DNA	3
Figure 1.3 Watson-Crick base pairs.....	4
Figure 1.4 The chain extension and termination reactions.....	7
Figure 1.5 Slab gel electrophoresis of the chain termination reaction products.....	9
Figure 1.6 The Maxam-Gilbert chemical degradation reactions.....	11
Figure 1.7 Slab gel electrophoresis of a Maxam-Gilbert sequencing sample	12
Figure 1.8 The peak height method developed by Richardson and Tabor.....	17
Figure 1.9	19
a. Polymerase chain reaction.....	19
b. Cycle sequencing.....	19
Figure 1.10	25
a. Random or shotgun sequencing	25
b. Primer walking	25
Figure 1.11 The succinylfluorescein dyes used by DuPont	30
Figure 1.12 The ABI fluorescent dyes.....	31
Figure 2.1	37
a. Schematic diagram of the capillary electrophoresis instrument.....	37
b. The sheath flow cuvette.....	38
Figure 2.2 Schematic diagram of the one-laser two-channel fluorescence detector	39
Figure 2.3 Electropherogram of an A-terminated M13mp18 sequencing sample	44

Figure 2.4 Retention time as a function of fragment length for single stranded DNA sequencing fragments	45
Figure 2.5 Peak spacing of DNA fragments for denaturing polyacrylamide gels.....	46
Figure 2.6 Sequencing rate for the data of Figure 2.4	48
Figure 2.7 Retention time for a vanishingly small DNA fragment as a function of total acrylamide concentration	49
Figure 2.8 Mobility of fluorescently labelled DNA fragments	52
Figure 2.9 Plate count for an 85 base DNA sequencing fragment as a function of total acrylamide concentration	54
Figure 2.10 Predicted plate count in a 4% total acrylamide gel	59
Figure 2.11 Resolution of bases 85-86	61
Figure 2.12 Predicted resolution for adjacent DNA fragments as a function of total acrylamide concentration	62
a. Fragments 250-251 bases in length	62
b. Fragments 500-501 bases in length	62
Figure 3.1 Diagram of the two-channel fluorescence detector.....	70
Figure 3.2.....	76
a. Electropherogram of A and C terminated MMTV DNA sequencing sample.....	76
b. The expanded electropherogram.....	78
c. Fragments 40 to 50 and 320 to 330 nucleotides in length	79
Figure 3.3	81
a. Electropherogram of the T and G terminated MMTV DNA sequencing sample.....	81
b. The expanded electropherogram.....	83
c. Fragments 20 to 30 and 110 to 120 nucleotides in length	84

Figure 3.4 Electropherogram of MMTV sample with a peak height ratio of 2:1.....	86
Figure 3.5 Electropherogram of an MMTV sample with a peak height ratio of 3:1.....	89
Figure 3.6 Electropherogram of an M13mp18 sample	91
Figure 3.7 Electropherogram of mouse cytokine DNA sample	93
Figure 4.1	105
a. Electropherogram of malaria DNA clone 1	105
b. Expansion showing fragments 100 to 150 and 350 to 400 nucleotides.....	107
Figure 4.2	109
a. Electropherogram of malaria DNA clone 2	109
b. Electropherogram of malaria DNA clone 2 - different peak height coding.....	111
Figure 4.3 Electropherogram of malaria DNA clone 3.....	113
Figure 4.4 Electropherogram of malaria DNA clone 4.....	115
Figure 4.5 Electropherogram of malaria DNA clone 5.....	117
Figure 4.6 Comparison of resolution in automated slab gel electrophoresis and CGE.....	122
Figure 5.1 The bases thymine and uracil.....	125
Figure 5.2 Diagram of the two-laser, two-channel fluorescence detector	129
Figure 5.3 Electropherogram of an A and C terminated M13mp18 sample labelled with F-12-dUTP.....	135
Figure 5.4 Sequencing Electropherogram for the two-colour internally labelled M13mp18 sample	139
Figure 5.5 Data processing to compensate for the mobility shift	140
Figure 5.6 Electropherogram of the internally labelled MMTV sample	142

Figure 5.7 Electropherogram of the internally labelled malaria sample	144
Figure 5.8 Electropherogram of the products of the labelling reaction	145
Figure 5.9	148
a. Electropherogram of the internally labelled C terminated M13mp18 sample.....	148
b. Electropherogram of the primer labelled C terminated M13mp18 sample.....	150
Figure 5.10 Retention time vs fragment length for internal and primer labelled samples.....	151
Figure 5.11	153
a. Inverse fragment length vs. mobility.....	153
b. A straight line fit to the data for fragments longer than 200 bases.....	153
Figure 5.12 Resolution as a function of fragment length for primer and internal labelled samples	154
Figure B.1 The effect of formamide on compressions.....	172
Figure B.2 Non-specific priming	175
Figure B.3 Template degradation	178
a. A C-terminated sample	178
b. An A-terminated sample.....	180

CHAPTER 1

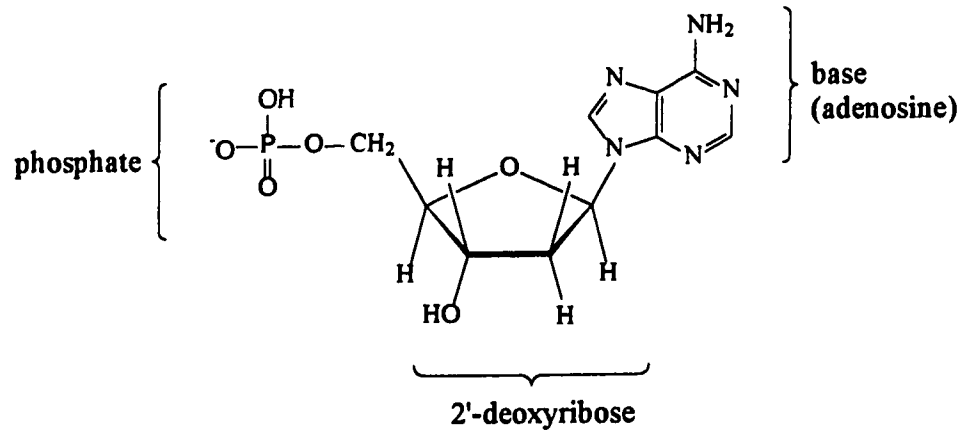
Introduction

Deoxyribonucleic acid (DNA) has been dubbed the blueprint for life. It is one of the most highly studied molecules, fascinating scientists for four decades. DNA was once described by Maurice Wilkins as "...Midas' gold. Everyone who touches it goes mad..." (1). DNA continues to fascinate. A current quest of scientists and organisations all over the world is the sequence of the DNA which makes up the human genome. On a smaller scale, the sequences of various genes are also sought. The very scale of such sequencing projects drives the continued development of new technologies. The subject of this thesis is the application of capillary gel electrophoresis with laser induced fluorescence detection to DNA sequencing.

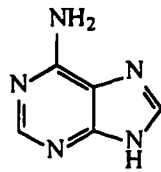
1.1 The Structure of DNA

DNA was first identified as the genetic principle in 1944 by Oswald Avery (2). By the year 1953, Watson and Crick had elucidated the structure of DNA (3). DNA is a long, straight chain polymer. The basic repeating unit of DNA is the nucleotide (Figure 1.1a). A nucleotide is composed of three substituents: 2'-deoxyribose, a phosphate group, and a purine or pyrimidine base. There are four different bases which may be attached to the ribose ring (Figure 1.1b), adenosine (A) and guanine (G), cytosine (C) and thymine (T). The nucleotides are joined together by a phosphodiester bond to form DNA. The 3' hydroxyl group of one nucleotide is joined to the 5' hydroxyl of the next. A strand of DNA has polarity: there is a 5' end with a free phosphate group, and 3' end with a free hydroxyl group. Single stranded DNA is shown in figure 1.2.

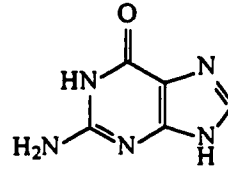
a.



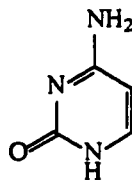
b.



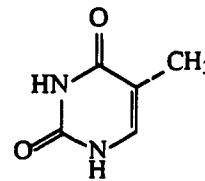
Adenosine



Guanine



Cytosine



Thymine

Figure 1.1

a. A nucleotide (deoxyadenosine monophosphate) is the basic repeating unit of DNA. The nucleotide is composed of three parts, a purine or pyrimidine base, in this case adenine, a 2'-deoxyribose, and a phosphate group.

b. The purine bases (top row), and the pyrimidine bases (bottom row)

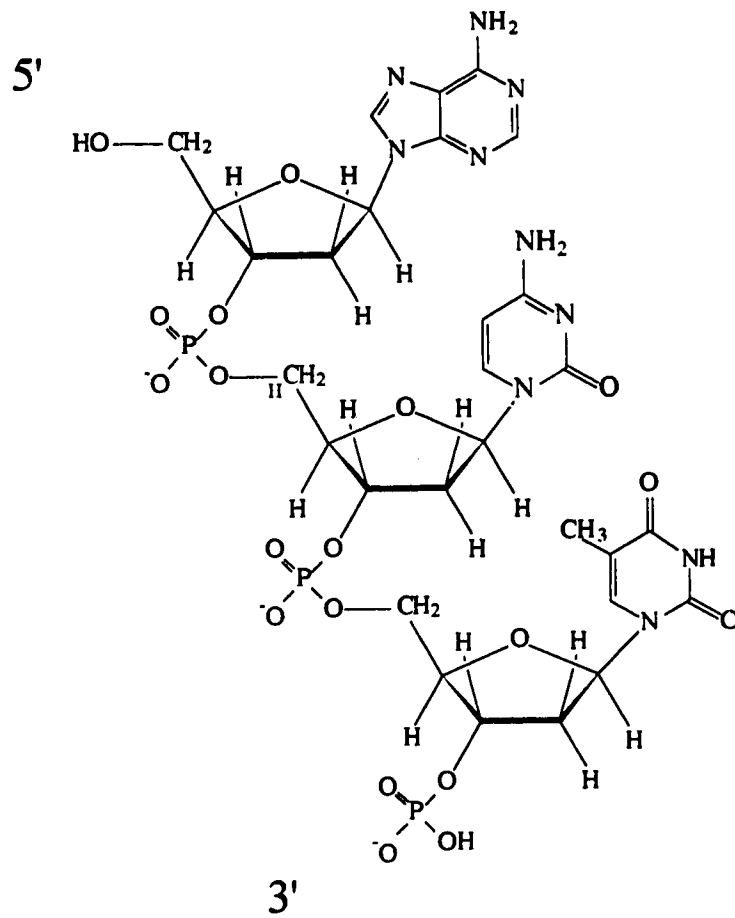


Figure 1.2 Single stranded DNA. This trinucleotide may be represented using the one letter codes for the bases as 5' ACT 3'.

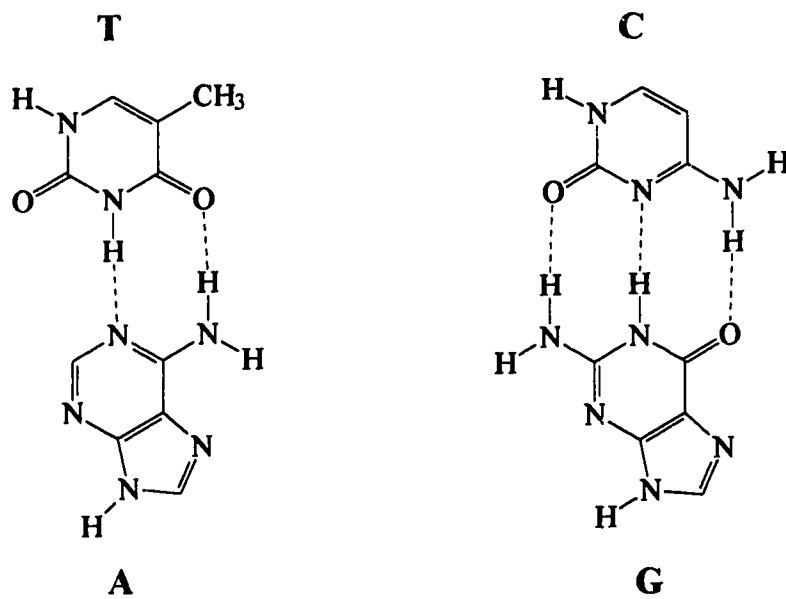


Figure 1.3 Watson-Crick base pairs. Adenine (A) pairs with thymine (T), on the left, and guanine(G) pairs with cytosine (C), on the right.

Generally DNA exists in the double stranded form, *i.e.* two DNA strands are associated. The two strands are held together by hydrogen bonds that form between the bases. The position of the hydrogen bonds makes pairing quite exclusive; A pairs with T and G pairs with C to form the Watson-Crick base pairs shown in Figure 1.3. The two strands of DNA are complementary by virtue of the base pairing: where there is A in one strand there is T in the other, where there is G in one strand there is C in the other. The two strands are also antiparallel; the 3' end of one strand lines up with the 5' end of the other strand.

The hydrogen bonds that hold the two strands together may be broken by raising the pH or the temperature of the DNA. In this way the DNA is denatured. Lowering pH or temperature means that the hydrogen bonds reform, *i.e.* the strands anneal. Short strands of DNA (oligonucleotides) which are complementary to a stretch of the DNA also easily anneal or hybridise.

1.2 DNA Sequencing Techniques

The sequence of DNA is the order of the nucleotides in a particular strand of DNA. The variable part of the DNA is the base attached to the sugar-phosphate backbone. The basic principle of most DNA sequencing techniques is the generation of a set of DNA fragments, which have one end in common and which differ in length by one nucleotide. In this way the sequence of nucleotides is converted into chain length information. What is needed to sequence the nucleotides is: 1) a way to generate the fragments, 2) a way to separate them on the basis of size, and 3) a way to determine the identity of the unique end of each fragment.

The most widely used method which generates a nested set of DNA fragments is the chain extension and termination method developed by Sanger (4). A second method based on chemical degradation of the DNA was developed by Maxam and Gilbert (5).

1.2.1 Chain Extension and Termination - Sanger's method

The chain extension and termination method developed by Sanger (4) makes use of the enzyme DNA polymerase. This enzyme requires a template DNA molecule of interest and a short oligonucleotide primer that is complementary to a known portion of the template. The polymerase synthesises a strand of DNA that is complementary to the template DNA by successive addition of the appropriate deoxynucleotide to the 3' hydroxyl group of the primer. The primer is annealed to the template, then the mixture is split into four tubes and DNA polymerase and the four deoxynucleotide triphosphates (dNTPs), are added to each tube. Finally, a small amount of 2'3'-dideoxyadenosine triphosphate (ddATP) is added to the first tube, dideoxycytidine triphosphate (ddCTP) to the second, dideoxyguanosine triphosphate (ddGTP) to the third, and dideoxythymidine triphosphate (ddTTP) to the fourth. Polymerase extends the primer to form a chain complementary to the template. Occasionally the polymerase incorporates a dideoxynucleotide triphosphate (ddNTP) that causes termination of the chain extension since the ddNTP lacks a 3' hydroxyl group. The product of the chain extension and termination reaction in the first tube is a set of DNA fragments that have a common 5' end, defined by the primer, and that all terminate with dideoxyadenosine (ddA). The second reaction produces fragments ending in dideoxycytidine, the third, dideoxyguanosine, and the fourth, dideoxythymidine. This process is presented schematically in Figure 1.4. Sanger

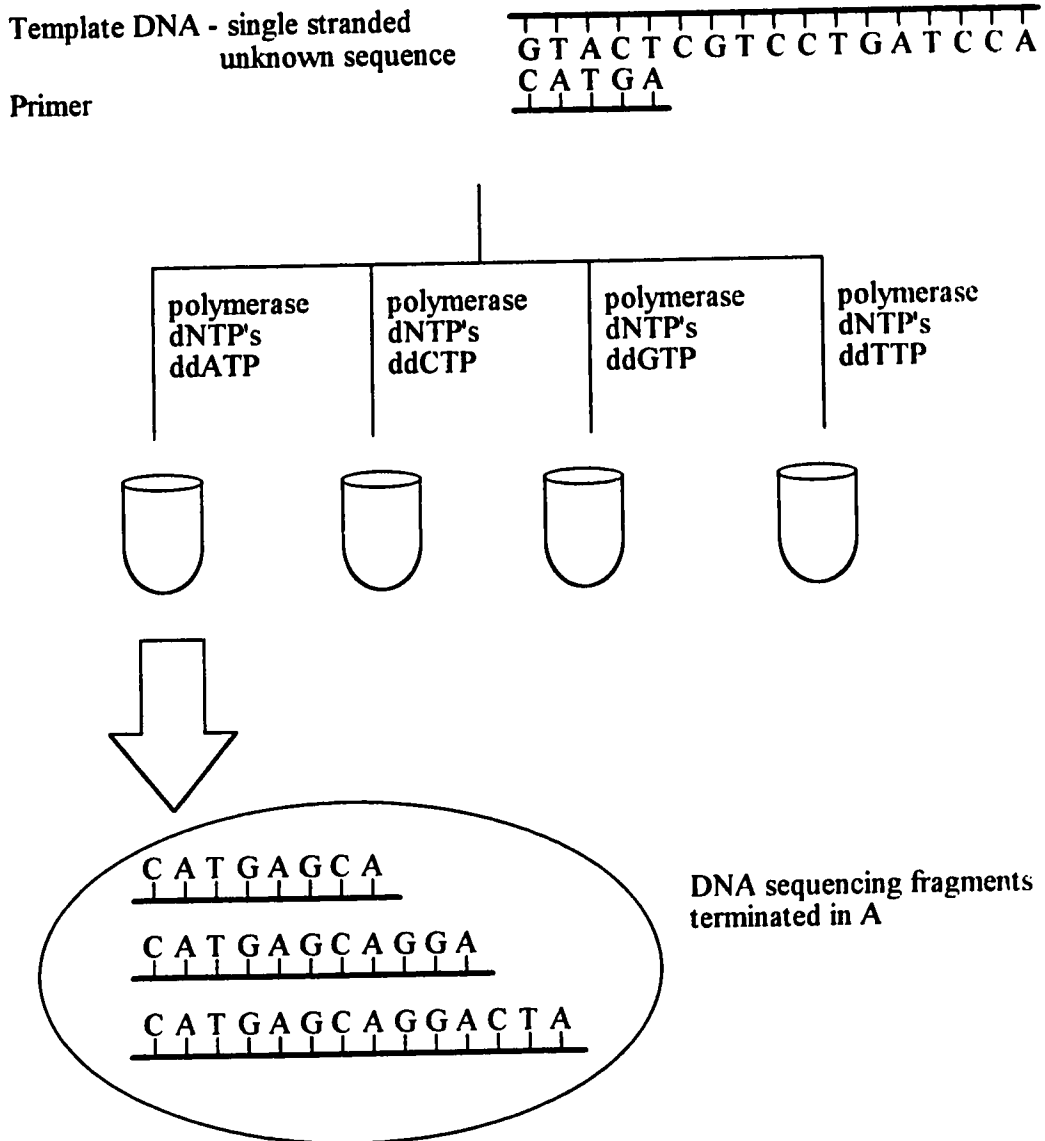


Figure 1.4 The chain extension and termination reactions - only the A sample is shown in detail.

labelled the DNA fragments by incorporation of a radioactively labelled deoxynucleotide in the synthesised DNA fragments. The DNA fragments are separated on the basis of size by polyacrylamide gel electrophoresis. The samples are loaded on four adjacent lanes of a rectangular slab of polyacrylamide. The application of an electric potential across the slab drives the separation. The negatively charged DNA fragments migrate down the gel. The smaller fragments migrate through the polymer network more quickly than the larger fragments. After a period of time the electrophoresis is stopped by turning off the electric potential. The bands of DNA are detected by exposing a film to the gel for a period of time, a process called autoradiography. The resulting autoradiogram gives the DNA sequence. The identity of the terminal nucleotide is indicated by the lane where the band is seen. The sequence is read from the bottom to the top of the autoradiogram (Figure 1.5).

1.2.2 The Maxam-Gilbert Chemical Degradation Technique

The Maxam-Gilbert technique for DNA sequencing relies on selective chemical degradation of the DNA to produce the nested set of fragments (5). The DNA is generally radiolabelled at one end to enable detection by autoradiography. It is then divided into five reaction tubes. The DNA in each tube is partially cleaved in a chemical reaction specific for one base e.g. G or type of base e.g. purine, A or G. The bases undergo chemical modification as summarised in Table 1.1, followed by cleavage of the sugar-phosphate backbone of the DNA by hot piperidine at the sites of chemical modification. The resulting cleavages occur for G, A+G, C+T, C and A>C. Partial

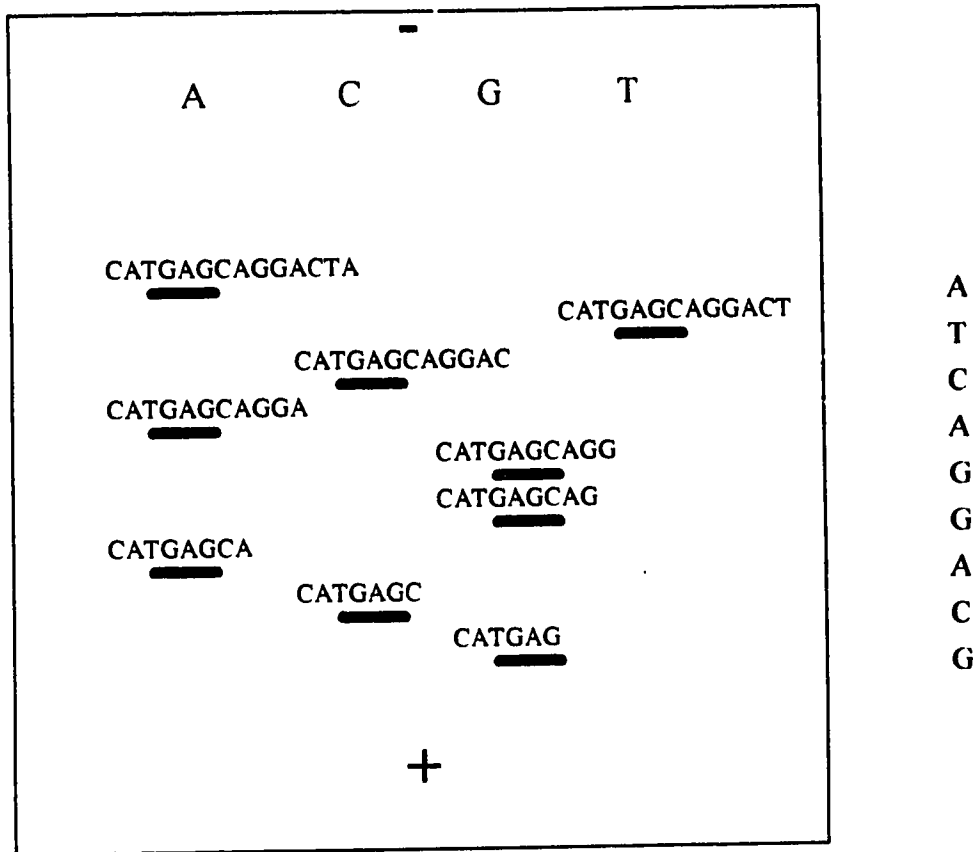


Figure 1.5 Slab gel electrophoresis. The four sequencing samples are loaded on adjacent lanes of the gel. The fragments migrate toward the positive electrode and are separated on the basis of size. After autoradiography, the film would appear similar to the diagram above, a series of dark bands. The corresponding DNA fragment is printed above each band for clarity. The identity of the 3' terminal nucleotide is determined from the lane where the band occurs. The sequence is read from bottom to top, since the shorter fragments migrate more quickly, hence move further toward the bottom of the gel.

cleavage yields a mixture of fragments, whose lengths are determined by the position of the base(s) in the original DNA (Figure 1.6).

Table 1.1 Chemical Modifications for Maxam-Gilbert Sequencing (6)

Base	Reagent	Resulting Modification
G	dimethyl sulphate, pH 8.0	Makes C8-C9 bond susceptible to base cleavage
A+G	piperidine formate, pH 2.0	protonates N in purines resulting in depurination
C+T	hydrazine	opens pyrimidine rings, they recyclise, and are susceptible to removal
C	1.5M NaCl, hydrazine	only C will react as above
A>C	1.2M NaOH	strong cleavage at A, weaker cleavage at C

The products of the reactions are loaded on five adjacent lanes of a gel, separated by electrophoresis, and detected by autoradiography. Comparison of the lanes on the gel yields the DNA sequence (Figure 1.7). The Maxam-Gilbert method generally gives sequences up to 250 nucleotides from the radiolabelled end of the DNA (6).

Note that the chain termination method yields the complementary sequence of the starting DNA, while the Maxam-Gilbert method gives the sequence of the starting DNA directly.

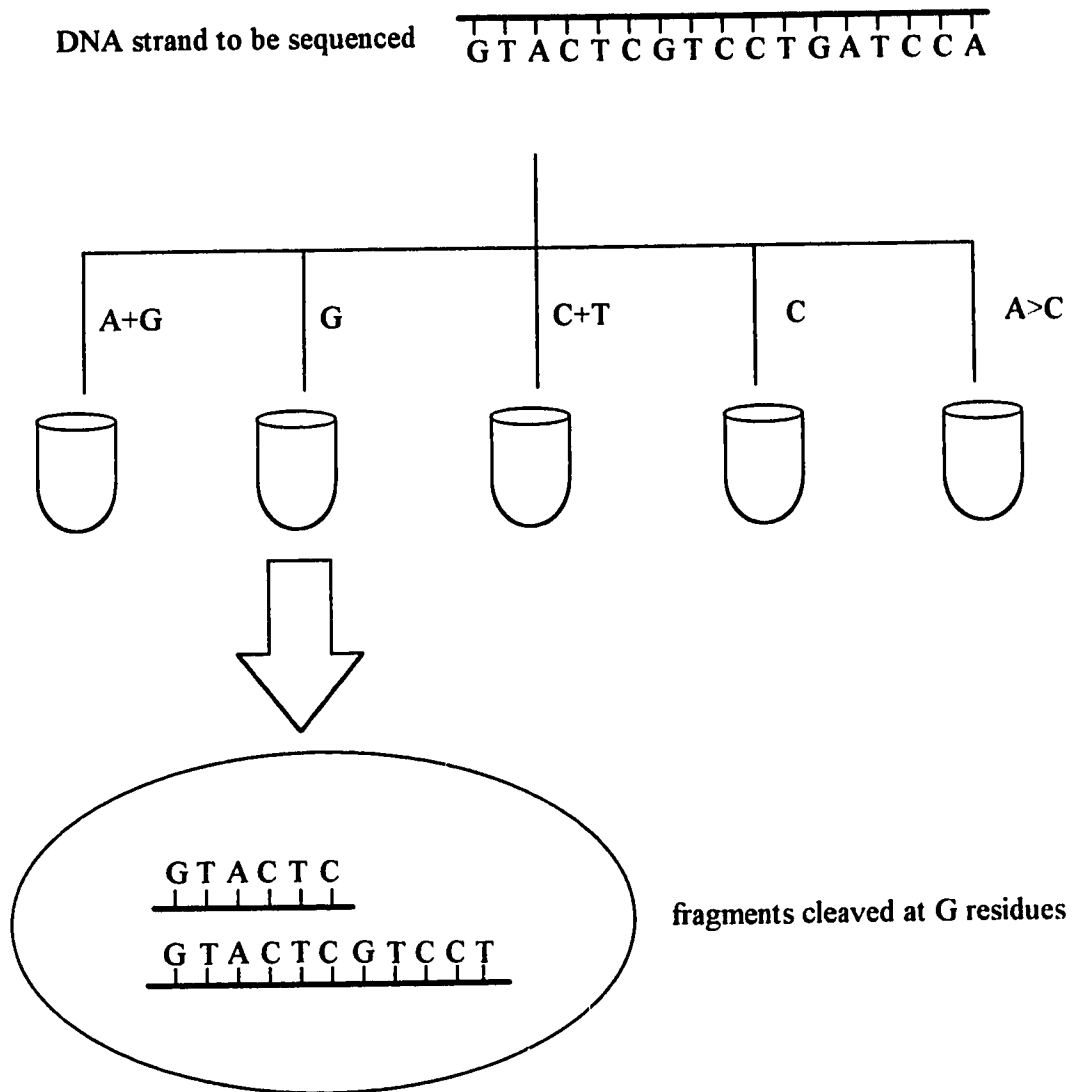


Figure 1.6 The Maxam-Gilbert chemical degradation reactions. The cleavage at the G residues are shown in detail.

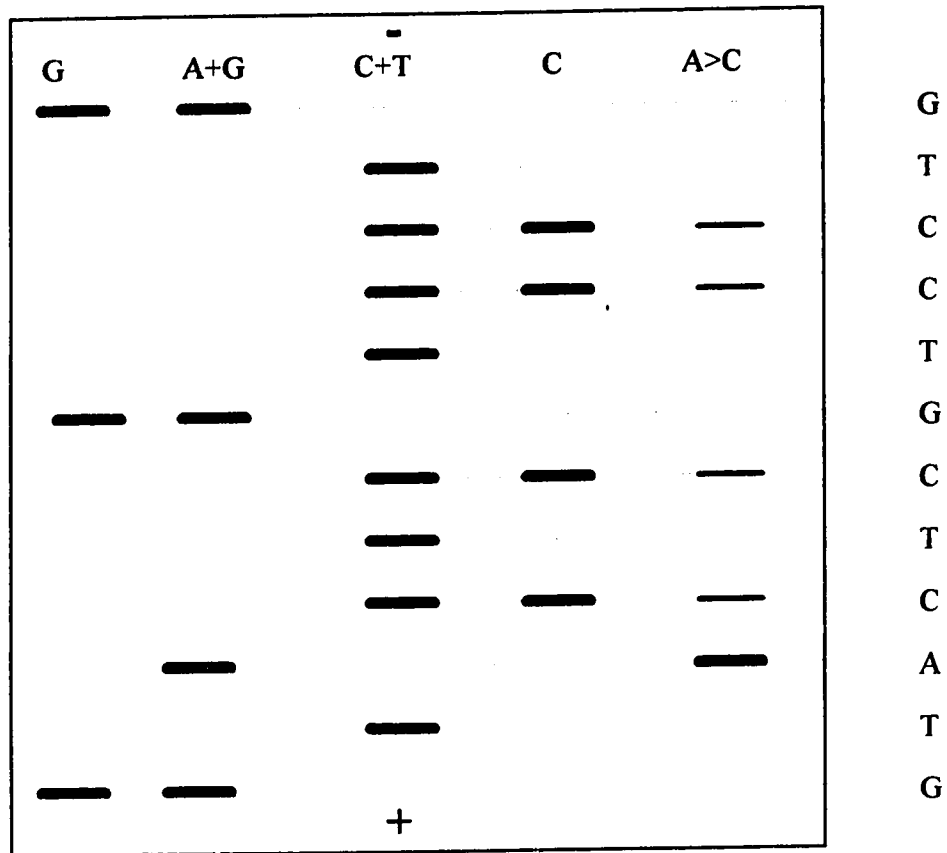


Figure 1.7 Slab gel electrophoresis of a Maxam-Gilbert sequencing sample. The DNA sequence is read from bottom to top. The identity of the base is established by comparison of the different lanes.

1.3 Modified Sequencing Approaches

While Sanger's method presented a major breakthrough in DNA sequencing, it is still labour intensive and relatively time consuming. Once the DNA samples have been prepared, approximately 8 to 16 hours is required for electrophoresis, 8 to 16 hours for autoradiography, and up to four hours to read and check the sequence. A typical sequencing run yields the sequence of 300 to 400 nucleotides. In addition to the time requirements, traditional sequencing methods require the use of radioactive nucleotides. In the mid 1980's the use of fluorescently labelled oligonucleotide primers was demonstrated by several different groups (7-9). This led to the development of automated DNA sequencers, which in a period of 8 to 12 hours perform electrophoresis, detect the fluorescent fragments, and determine the nucleotide sequence (10). The following sections provide an overview of the most widely implemented approaches to fluorescent labelling of DNA samples and the implications for instrument design.

1.3.1 One Fluorescent Primer

The simplest approach to fluorescent labelling is to incorporate a fluorescent label on the 5' end of the primer. This approach was demonstrated by Ansorge in 1986 (7). All samples have the same label, hence four separate sequencing reactions are carried out, and the products of the reactions are separated on four adjacent lanes of a slab gel. The identity of the terminal nucleotide is determined by the lane where the band is located, analogous to autoradiography. The hallmark of this method is its simplicity; only one label is used, simplifying detector optics since only one laser is required to excite the fluorescent label and only one detector channel is necessary.

The disadvantages are that four separate reactions are required, and four lanes on the gel are used for only one sample. Both Pharmacia and Hitachi have developed automated sequencers based on this approach.

1.3.2 Four Fluorescent Primers

The approach pioneered by Smith *et al.* (8). makes use of four spectrally distinct fluorescent labels. In this approach, the four individual sequencing reactions are carried out. Each reaction makes use of a primer with one of the four labels. The reaction products are then pooled and separated on a single lane of a slab gel. The detector optics for this system are more complicated. Two lasers are required to excite all four labels, and four detector channels are necessary. Four separate sequencing samples are still required, but only one lane on the gel is used. The use of four distinct labels minimizes ambiguities. The Applied Biosystems automated DNA sequencer is based on this approach.

1.3.3 Dye terminators

The use of fluorescent ddNTP's or dye terminators was first demonstrated by Prober *et al.* at DuPont (11). Each of the four ddNTPs is labeled with a fluorescent dye, and all four ddNTPs were added to one DNA sequencing reaction. The four dyes are derivatives of fluorescein with closely spaced excitation and emission spectra. All four dyes are excited by an argon ion laser operating at 488 nm. The fluorescence is split into two spectral channels and the identification of the terminal nucleotide was based on the ratio of the signals in the two channels. The advantages of this approach are first, that only one sequencing reaction is required as well as only one lane on the

gel, and second, that false terminations (those not caused by the incorporation of a ddNTP) are not detected. Only one laser, and two spectral channels are required for the detector. Unfortunately this method suffers from poor accuracy due to the close spacing of the fluorescence spectra. The instrument based on this approach was manufactured by DuPont, but is no longer available.

The same principle has been applied with four fluorescent dyes that have more widely spaced spectra, and which are compatible with the ABI sequencer (12). These dyes are excited by two different laser lines, and are detected in four different spectral channels. This approach maintains the advantage of one reaction and one lane on the gel combined with the more complicated detector design.

1.3.4 Internal Labelling with Fluorescent dNTPs

In 1992 Voss *et al.* first reported labelling of DNA samples with a fluorescein tagged dUTP (13). The fluorescein-dUTP was incorporated in a separate labelling step similar to that used to incorporate radioactive nucleotides. In a separate report, labelling with fluorescein-dATP was reported by the same group in 1993 (14). Samples produced in this manner are all tagged with the same label, hence four reactions and four lanes on the gel are required. The samples are suitable for separation on the Pharmacia sequencer. The main advantage of internal labelling is the avoidance of the costly and time consuming preparation of labelled primers. The labelled dNTPs are incorporated more efficiently than the labelled ddNTPs by the DNA polymerases used in sequencing.

1.3.5 Peak Height Methods

In 1990 Richardson and Tabor demonstrated a one label, peak height encoded method for DNA sequencing (15). The basis of this method is the very uniform pattern of terminations produced by T7 polymerase in the presence of manganese ions (16). Fluorescent primers are used to label the samples, and the reaction is carried out in one tube. The dideoxynucleotides are added in different amounts, for example with a ratio of 8:4:2:1 of ddATP:ddGTP:ddCTP:ddTTP. In this way, the proportion of fragments terminated by ddA is greatest and by ddC is the least. The DNA sequence is encoded in the intensity of the signal; the most intense signal corresponds to A, the least intense, T, and the intermediate levels, G and C (Figure 1.8). This approach requires only one sample preparation, one lane on the gel, and a simple detector design since only one label is used. These samples may also be internally labelled with fluorescent dNTPs (Sec. 1.3.4). The main drawback is that the accuracy is not as high as for other methods. Two types of errors occur: 1. ambiguities in the two bases coded by the intermediate peak heights, and 2. late in the run as the peak spacing decreases, and as the least intense peaks become very small they may be missed completely.

1.3.6 Cycle Sequencing

Cycle sequencing(17-18) is an offshoot of the polymerase chain reaction (PCR) (19). PCR results in an exponential amplification of a segment of DNA by repetitive cycling (Figure 1.9a). Each cycle consists of three steps: 1. the high temperature denaturation of double stranded DNA, 2. the annealing of two complementary primers that flank the segment of DNA to be amplified, and 3. the extension of the

Template DNA - single stranded
 unknown sequence
 Primer with fluorescent label

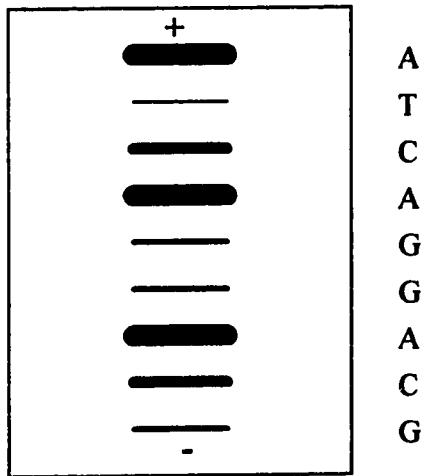
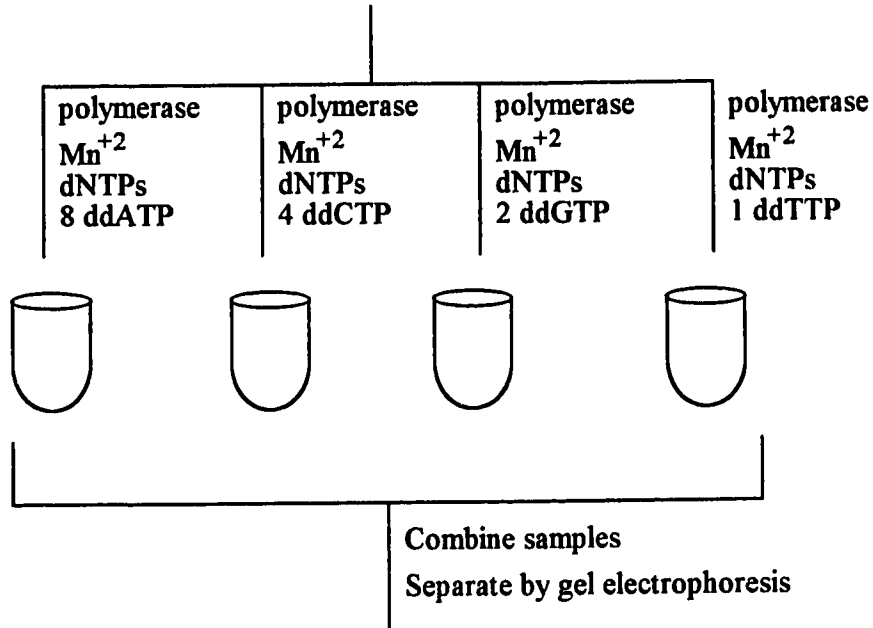


Figure 1.8 The peak height method developed by Richardson and Tabor. The ddNTP's are added in the ratio 8:4:2:1 A:C:G:T. The samples are combined and separated on a single gel lane. The intensity of each band identifies the 3' terminal nucleotide: The most intense bands are A's, the least intense are T's.

primers by DNA polymerase. The use of thermostable *Taq* polymerase, which withstands the denaturation step, was a significant advance (20). PCR allows very small amounts of DNA to be amplified to easily detectable amounts.

Cycle sequencing uses only one primer, to linearly amplify the target DNA, in the presence of ddNTPs that terminate each newly extended chain (Figure 1.9b). The amplified DNA is a nested set of fragments like those produced by Sanger's method (4). Amounts as low as nanograms of DNA template can be sequenced. Conventional sequencing reactions require microgram quantities of template DNA. In theory, cycle sequencing would avoid the time consuming preparation of large amounts of template.

1.4 Mapping and Sequencing the Genome

The human genome contains approximately 3×10^9 base pairs divided into 23 chromosomes. The genome also contains about 100 000 genes, the DNA sequences that code for proteins as well as information that regulates the kind and amount of protein made in a particular cell. The genes account for about 5% of the total DNA present in the genome, and they are interspersed with regions of non-coding DNA.

The goal of the Human Genome Project is to map and sequence the entire human genome within 15 years. A brief overview of some of the techniques used, and the steps necessary to get to DNA sequence starting from chromosomal DNA are presented below.

1.4.1 Restriction Digests

DNA molecules may be cut into fragments by restriction enzymes. Restriction enzymes are isolated from various types of bacteria. Each restriction enzyme cuts

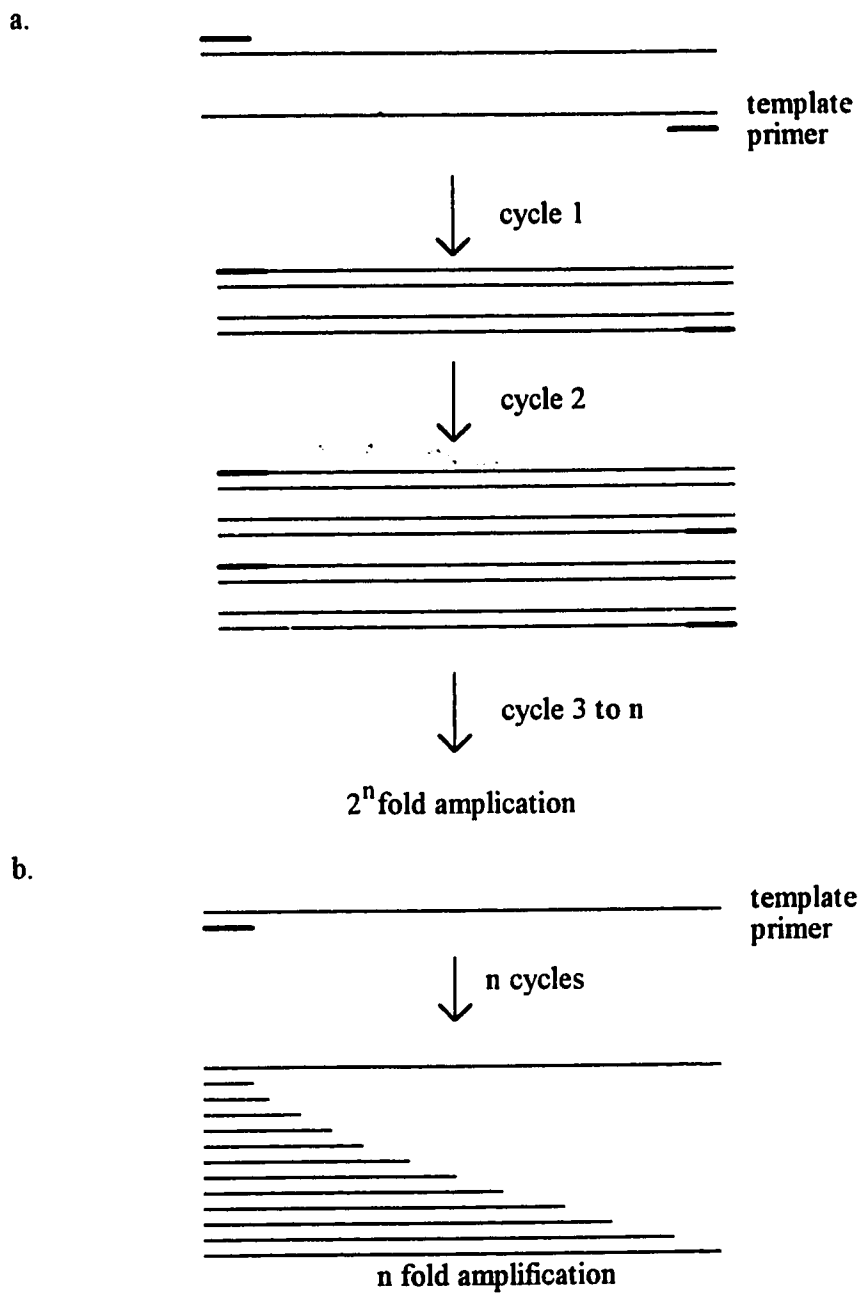


Figure 1.9 a. Polymerase chain reaction (PCR). Each successive cycle produces a twofold amplification of the target DNA.

b. Cycle sequencing. Each cycle produces a ddNTP terminated fragment resulting in a linear amplification and a nested set of fragments.

double stranded DNA at or adjacent to a unique recognition site. The enzyme recognition site may be 4 to 8 bases long. Some examples of restriction enzymes and their recognition sites are shown in Table 1.2 (21).

Table 1.2 Some restriction enzymes and their recognition sites. The cutting sites are indicated by the arrows.

Source	Enzyme	Recognition Sequence
<i>Haemophilus aegyptius</i>	<i>HaeIII</i>	GG↓CC CC↑GG
<i>Escherichia coli</i> RY13	<i>EcoRI</i>	G↓AATTC CTTAA↑G
<i>Bacillus amyloliquefaciens</i> H	<i>BamHI</i>	G↓GATCC CCTAG↑G
<i>Nocardia otitidis-caviarum</i>	<i>NotI</i>	GC↓GGCCGC CGCCGG↑CG

If DNA is a random sequence of the four deoxynucleotides, then an enzyme with a 4 base recognition site will give, on average, fragments 256 nucleotides (or bases) in length, one with a 6 base recognition site will give 4000 base fragments, and one with an 8 base recognition site will give 64 000 base fragments. Hundreds of different restriction enzymes are known. The resulting restriction fragments are useful for mapping and cloning experiments.

1.4.2 Gel Electrophoresis

DNA fragments may be separated on the basis of size by gel electrophoresis. Fragments of up to 1000 nucleotides can be separated with polyacrylamide gels. Agarose gel will separate fragments ranging from 100 to 50 000 nucleotides, depending on the gel concentration (21). Longer fragments may be separated using

pulsed field gel electrophoresis. Individual bands may be recovered from the gel and purified for further experiments.

1.4.3. Cloning

DNA fragments may be inserted into a circular DNA molecule known as a cloning vector to produce a recombinant DNA molecule. The vector is introduced into a bacterial cell where it is multiplied. The copies of the vector which are produced all contain the inserted DNA fragment. When the bacterial host cell multiplies the recombinant DNA is passed on to the new cells. As cell division proceeds a large number of identical host cells results, each containing multiple copies of the recombinant DNA molecule. Suitable vectors include plasmids, bacteriophages (phages), cosmids, and yeast artificial chromosomes (YACs). Plasmids are small circles of DNA that are found in bacteria and can replicate independently of the bacterial chromosome. A bacteriophage is a virus that infects bacterial cells and is replicated inside the cells. Plasmid and phages hold DNA inserts of up to about 10 kilobases (kb). A cosmid is a hybrid between a phage and a plasmid that will hold inserts of up to 40 kb in size. YACs are composed of the structural components of the yeast chromosome. YACs can hold very large inserts. (22-23)

A series of clones is generally prepared from the DNA of interest, for example chromosomal DNA. The collection of overlapping clones that represents the entire length of the starting DNA is called a library.

1.4.4 Mapping and Sequencing

Mapping involves breaking the DNA down into fragments small enough to

characterise. These fragments are ordered, or mapped to correspond to their approximate positions on the chromosome. At the coarsest level is the chromosome map, where genes or fragments of DNA are assigned to one of the chromosomes. Generally this is accomplished by hybridisation techniques: a piece of tagged DNA (either fluorescent or radioactive) that is complementary to a region of the gene being mapped is used to find and bind to the complementary strand of DNA in the chromosome. Chromosome maps typically locate a DNA fragment within a region of about 10 million base pairs (24).

Genetic maps are constructed by observation of the pattern of inheritance of pairs of genetic markers. A useful genetic marker is any molecular or physical characteristic that is inherited and that differs between individuals. If two markers are located close to each other on the chromosome, they tend to be passed together from parent to child. Occasionally, DNA strands break and rejoin, an event known as recombination. The closer two markers are, the less likely they are to be separated by a recombination event. Currently, genetic maps have a resolution of about 10 million base pairs. The strength of a genetic map is that inherited diseases can be located on the map, aiding in finding disease genes.

Restriction maps are physical maps, ordered sets of DNA fragments, that are prepared by cutting the genomic DNA with restriction enzymes. The map shows the positions of different restriction sites on a DNA molecule. A chromosome may be cut into small pieces with restriction enzymes. Each of the pieces is cloned and the clones are arranged in serial fashion to give a contiguous map known as a contig. Alternatively, the chromosome may be cut with rare-cutter restriction enzymes to give large pieces, which are cloned in YACs or other vectors, and are ordered. The fragments are then subdivided and mapped further. This approach gives a more continuous map, but coarser resolution.

Once the map is constructed, each of the vectors is sub-cloned. Sub-cloning is carried out because YACs are too large to be sequenced directly. The inserts in the YACs are mapped, divided into smaller pieces, and cloned in a different vector, such as a cosmid. The cosmids are further subdivided and grown in plasmids. The plasmids, for example pBluescript, are the templates used for DNA sequencing and may be produced in the single stranded form. Generally, one of two sequencing strategies is employed. The first approach is random or shotgun sequencing. In this approach a large number of subclones are prepared containing random segments of target DNA (Figure 1.10a). Once the sequence data for the fragments is obtained, a computer orders the final sequence. In general, it is necessary to sequence 5 to 7 times the length of the target DNA to generate the entire sequence (6). The second approach is a directed approach, where the sequences of the target DNA are obtained in a systematic way. An example of a directed approach is primer walking which sequences stretches of DNA in a stepwise fashion (Figure 1.10b). The sequence from the first reaction is used to design the primer for the next reaction. The DNA sequence of the entire insert is obtained by moving the priming site along the DNA (6).

The raw sequence data is assembled to give a finished sequence which is then edited.

1.5 Capillary Electrophoresis

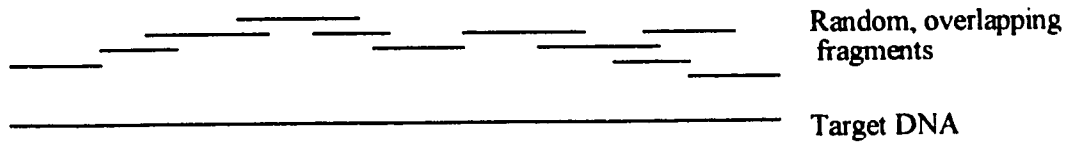
Electrophoresis is the separation of charged molecules based on their mobility in an applied electric field. Historically, electrophoresis has found widespread application in the separation of biological macromolecules such as proteins and nucleic acids. The efficiency of electrophoretic separations is limited by Joule heating which

causes convection. Usually electrophoresis is carried out with a solid support, *e.g.* paper, or with an anticonvective medium such as polyacrylamide gel. Hjerten first described open tubular zone electrophoresis (25). In 1983 Jorgenson and Lukacs described capillary electrophoresis (26), which has rapidly grown into a relatively mature field.

There are several advantages associated with capillary electrophoresis, which are related to the dimensions of the capillary. The small inner diameter of the capillaries, typically 10 to 100 μm , minimizes convection (27), removing the need for an anticonvective medium. The high surface-to-volume ratio dissipates heat efficiently, which allows application of high electric fields resulting in very fast and efficient separations. Separation efficiencies of up to 2.5 million theoretical plates are not uncommon in free zone capillary electrophoresis

The theory of open tubular capillary electrophoresis shows that the separation is based on the charge-to-radius ratios of the analytes. Application of free zone electrophoresis to separations of DNA is limited since the charge-to-radius ratio of DNA remains approximately constant as the chain size increases. A direct technology transfer was the key to separating DNA fragments by capillary electrophoresis. In 1987 Karger described the preparation of polyacrylamide gel filled capillary tubes (28). Swerdlow and Gesteland applied the gel filled capillaries to the separation of DNA sequencing samples (29). The gel acts as a sieving medium and separates the fragments on the basis of size. Traditionally, slab gel electrophoresis was carried out in crosslinked polyacrylamide gel. The composition of such gel is expressed as total

a.



b.

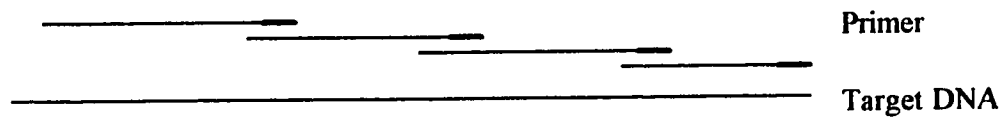


Figure 1.10 a. Random or shotgun sequencing. A series of random clones is generated and sequenced. The sequences of the fragments are ordered by a computer.

b. Primer walking. A portion of the target DNA is sequenced starting with a primer (shown by the short, heavy line). The newly obtained sequence is used to prepare a second primer, which is used to sequence the next portion of the target DNA. This is continued until the target DNA is sequenced.

percent acrylamide, %T, and percent crosslinker, %C.

$$\%T = \frac{W_{\text{Acrylamide}}(\text{g}) + W_{\text{Bis}}(\text{g})}{V(\text{ml})} \times 100\% \quad (1-1)$$

$$\%C = \frac{W_{\text{Bis}}(\text{g})}{W_{\text{Acrylamide}}(\text{g}) + W_{\text{Bis}}(\text{g})} \times 100\% \quad (1-2)$$

where the weights of acrylamide, $W_{\text{Acrylamide}}$, and the crosslinker N,N'-methylenebisacrylamide (Bis), W_{Bis} , are in grams, and the total volume of the solution is in millilitres. The mobility, μ , of a DNA fragment of a given length is sometimes described by a function of the gel concentration:

$$\log \mu = \log \mu_0 + k(\%T) \quad (1-3)$$

where μ_0 is the mobility in free solution (%T=0), k is the retardation coefficient, and %T is the total acrylamide concentration, where crosslinker concentration, %C, is held constant. High resolution separations of oligonucleotides are also carried out in non-crosslinked polyacrylamide (30).

1.6 Laser-Induced Fluorescence

The impressive speed and efficiency of separations performed by capillary electrophoresis are directly linked to the small dimensions of the capillary. Ironically, the capillary's dimensions and shape also produce the biggest challenge in implementing capillary electrophoresis, that of detection. A capillary, typically with dimensions of 50 μm i.d. and 50 cm long, has a volume of less than one microlitre. The concentration of analyte must be appreciably lower than the ionic strength of the

separation buffer in order to produce efficient, reproducible separations. As a result the volume of sample injected into the capillary is on the order of one nanolitre or less. Very high sensitivity detectors are required and include laser induced fluorescence (31), laser induced thermo-optical absorbance (32), electrochemistry (33), UV absorbance (34), radioactive isotopes (35), and mass spectrometry (36). The sensitivity of laser-induced fluorescence is sufficient to allow detection of very small amounts of samples.

The fluorescence intensity, I , is expressed as

$$I = QI_0(1 - 10^{-abc}) \quad (1-4)$$

where Q is the quantum yield of the fluorophore, I_0 is the incident power of the excitation source, a is the molar absorptivity of the molecule, b is the sample path length, and c is the sample concentration. If the sample concentration is sufficiently low the following approximation is true

$$I = QI_0abc \quad (1-5)$$

Equation 1-5 is further modified to

$$I = KQI_0abc \quad (1-6)$$

Equation 1-6 includes an instrumental factor K , where $K < 1$, reflecting the fact that fluorescence is emitted in all directions but is only viewed within a limited aperture. Capillary separations require that the sample concentration, c , and the path length, b , are small. The sensitivity may be increased by choice of a molecule with high quantum

yield Q and high molar absorptivity a , by matching the excitation source to the absorption maximum, increasing the source intensity I_0 , and increasing the collection efficiency, which is related to K .

Laser induced fluorescence is ideal for detection in capillary electrophoresis. The spatial coherence of the laser beam allows it to be very tightly focussed to illuminate very small sample volumes with high intensity, I_0 . When coupled with efficient collection and detection, very small volume samples may be detected.

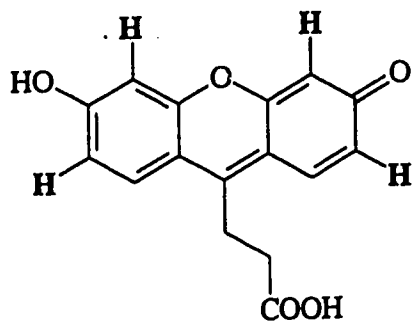
There are several sources of background signal in fluorescence measurements: fluorescence from impurities in the solvent, fluorescence from cuvette windows, Raman and Rayleigh scatter from the solvent, light scatter at the cuvette-sample interface (37) and detector dark current. Careful choice of reagents and sample preparation minimize background from the first source. Solvent Raman scatter is rejected by appropriate spectral filtering. Fluorescence and light scatter from the cuvette windows are minimized by spatial filtering, which restricts the field of view of the detector to the illuminated sample stream. If fluorescence detection in capillaries is performed on-column, light scatter from the capillary window is the major source of background signal. Use of the sheath flow cuvette as a post-column detector eliminates light scatter from the capillary interface, and lowers the background signal to nearly Raman scatter or even dark current limited values (38-39).

Most samples are not fluorescent and must be tagged with an appropriate label in order to be detected. Some popular labels are fluorescein and tetramethylrhodamine. Fluorescein has an absorbance maximum at 494 nm and is efficiently excited by the argon ion laser line at 488 nm. Tetramethylrhodamine, with an absorbance maximum at 547 nm, is excited by the green helium-neon laser line at 543.5 nm. A variety of samples have been labelled using these compounds, separated by capillary electrophoresis and detected by laser-induced fluorescence. Samples

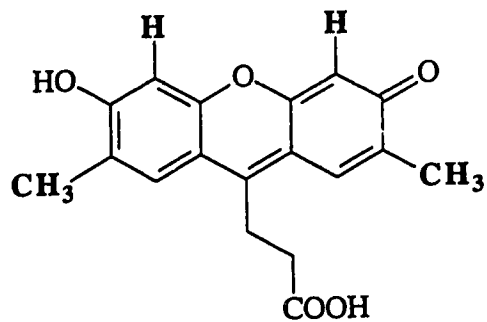
include amino acids (31), peptides (40), monosaccharides (41), and oligonucleotides (39).

A variety of labels have been used for the detection of DNA sequencing samples. The one-colour approach (sec 1.3.1) generally uses fluorescein (Figure 1.12). The dye terminator approach developed by DuPont (sec 1.3.3) used several modified fluorescein compounds (Figure 1.11). The dye terminator approach developed by ABI uses four different dyes with more widely spaced spectra. The four colour labelled primer approach (sec. 1.3.2) also uses four distinct dyes (Figure 1.12). The work described in this thesis uses a variety of labels. The DuPont dye terminators are used in Chapter 2, two of the ABI dye primers (FAM and JOE) in Chapters 3 and 4, and fluorescein and tetramethylrhodamine, in Chapter 5.

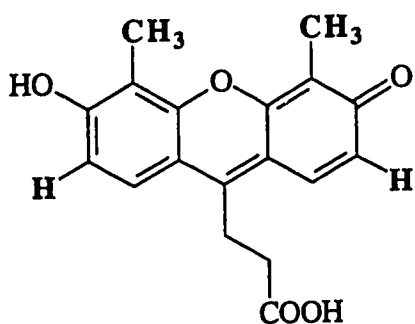
This thesis describes the effect of acrylamide concentration, %T, on the separation of DNA fragments by capillary electrophoresis, and describes a variety of approaches to DNA sequencing by capillary electrophoresis with laser induced fluorescence detection.



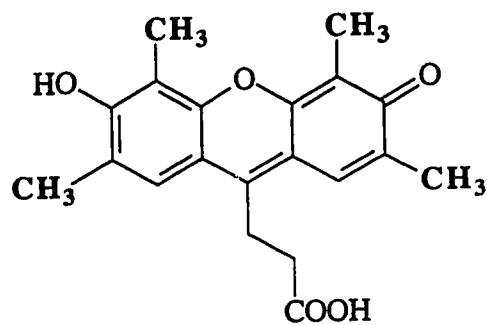
SF-505



SF-512

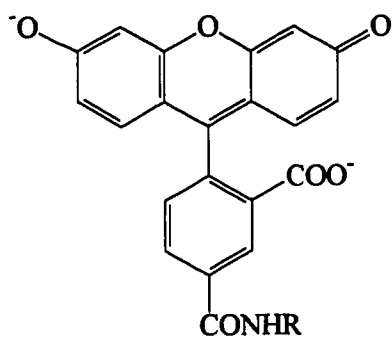


SF-519



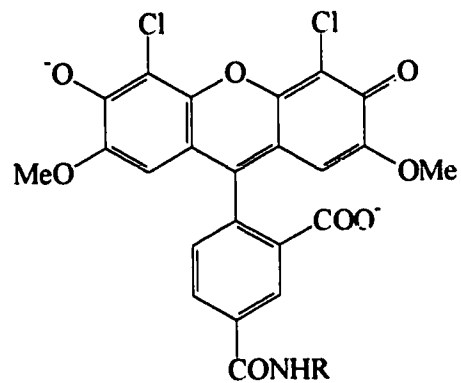
SF-526

Figure 1.11 The succinylfluorescein dyes used in the DuPont dye terminators. The number under each structure indicates the position (nm) of the emission maximum.



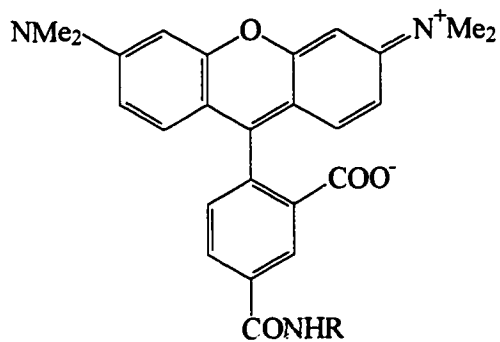
FAM (fluorescein)

Abs max = 494nm
Em max = 520nm



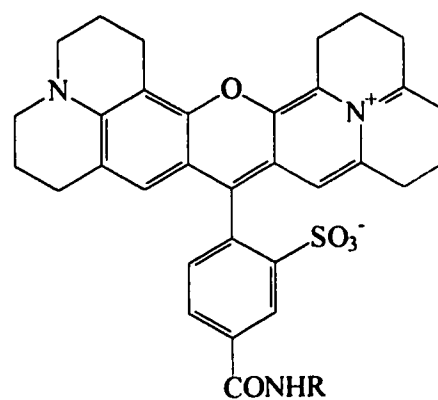
JOE

Abs max = 524nm
Em max = 550nm



TAMRA (tetramethylrhodamine)

abs max = 547nm
em max = 572nm



ROX

abs max = 589nm
em max = 615 nm

Figure 1.12 Fluorescent dyes used in the ABI dye primers.

References

1. H.F. Judson, *The Eighth Day of Creation*, New York, Simon & Schuster, 1979.
2. O.T. Avery, C.M. MacLeod, M. McCarty, *J. Exptl. Med.*, **79** (1944), 137.
3. J.D. Watson, F.H.C. Crick, *Nature*, **171** (1953), 737.
4. F. Sanger, S. Nicklen, A.R. Coulson, *Proc. Natl. Acad. Sci. USA*, **74** (1977), 5463.
5. A.M. Maxam, W. Gilbert, *Proc. Natl. Acad. Sci.*, **74** (1977), 560.
6. J. Sambrook, E.F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual*, Second Ed., Plainview, NY, Cold Spring Harbor Laboratory Press, 1989.
7. W. Ansorge, B.S. Sproat, J. Stegemann, C. Schwager, *J. Biochem. Biophys. Meth.*, **13** (1986), 315
8. L.M. Smith, J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connell, S.B.H. Kent, L.M. Hood, *Nature*, **321** (1986), 674.
9. H. Kambara, T. Nishikawa, T. Katayama, T. Yamaguchi, *BioTechnology*, **6** (1988), 816.
10. T. Hunkapillar, R.J. Kaiser, B.F. Koop, L. Hood, *Science*, **254** (1991), 59.
11. J.M. Prober, G.L. Trainor, R.J. Dam, F.W., Hobbs, C.W. Robertson, R.J. Zagursky, A.J. Cocuzza, M.A. Jensen, K.R. Bauermeister, *Science*, **238** (1987), 336.
12. L.G. Lee, C.R. Connell, S.L. Woo, R.D. Cheng, B.F. McArdle, C.W. Fuller, N.D. Halloran, R.K. Wilson, *Nucl. Acids Res.*, **20** (1992), 2471.
13. H. Voss, C. Schwager, U. Wirkner, J. Zimmermann, H. Erfle, N.A. Hewitt, T. Rupp, J. Stegemann, W. Ansorge, *Meth. Molec. Cell. Biol.*, **3** (1992), 30.

14. H. Voss, S. Wiemann, U. Wirkner, C. Schwager, J. Zimmermann, J. Stegemann, H. Erfle, N.A. Hewitt, T. Rupp, W. Ansorge, *Meth. Molec. Cell. Biol.*, **3** (1992), 153.
15. S. Tabor, C.C. Richardson, *J. Biol. Chem.*, **265** (1990), 8322.
16. S. Tabor, C.C. Richardson, *Proc. Natl. Acad. Sci. USA*, **86** (1989), 4076.
17. A.M. Carother, G. Urlaub, J. Mucha, D. Grunberger L.A. Chasin, *BioTechniques*, **7** (1989), 494.
18. V. Murray, *Nucl. Acids Res.*, **17** (1989), 8889.
19. K. Mullis, F.A. Faloona, *Methods Enzymol.*, **155** (1987), 335.
20. R.K. Saiki, D.H. Gelfand, S. Stoffel, S.J. Scharf, R. Higuchi, G.T. Horn, K.B. Mullis, *Science*, **239** (1988), 487.
21. D.A. Micklos, G.A. Freyer, *DNA Science: A First Course in Recombinant DNA Technology*, 1990, Cold Spring Harbor Laboratory Press, .
22. T.A. Brown, *Gene Cloning: An Introduction*, Second Ed., London, Chapman & Hall, 1990.
23. T.A. Brown, *Genetics: A Molecular Approach*, Second Ed., London, Chapman & Hall, 1992.
24. C. Cantor, S. Spengler, *Primer on Molecular Genetics*, published in DOE *Human Genome 1989-90 Program Report*, revised and expanded by D. Casey, modified for Web access by D. Jacobson. Mosaic:
<http://wwwwgdb.org/Dan/DOE/intro.html>
25. S. Hjerten, *Chromatogr. Rev.*, **9** (1967), 122.
26. J.W. Jorgenson, K.D. Lukacs, *Science*, **222** (1983), 266.
27. F.E.P. Mikkers, F.M. Everaerts, J.P.E.M. Verheggen, *J. Chromatogr.*, **169**, 11-20 (1979).
28. A.S. Cohen, B.L. Karger, *J. Chromatogr.*, **397** (1987), 409.

29. H. Swerdlow, R. Gesteland, *Nucl. Acids Res.*, **18** (1990), 1415.
30. M.C. Ruiz-Martinez, J. Berka, A. Belenkii, F. Foret, A.W. Miller, B.L. Karger, *Anal. Chem.*, **65** (1993), 2851.
31. Y.F. Cheng, N.J. Dovichi, *Science*, **242** (1988), 562.
32. M.Yu, N.J. Dovichi, *Appl. Spectrosc.*, **43** (1989), 196.
33. R.A. Wallingford, A.G. Ewing, *Anal. Chem.*, **60** (1988), 1973.
34. Y. Walbroehl, J.W. Jorgenson, *J. Chromatogr.*, **315** (1984), 135.
35. S.L. Pentoney, R. Zare, J. Quint, *J. Chromatogr.*, (1989), 480, 259-270.
36. R.D. Smith, J.A. Olivares, N.T. Nguyen, H.R. Usdeth, *Anal. Chem.*, **60** (1988), 436.
37. C.A. Parker, *Photoluminescence of Solutions*, New York, Elsevier, 1968, pp. 411-426.
38. S. Wu, N.J. Dovichi, *J. Chromatogr.*, **480** (1989), 141.
39. D.Y. Chen, H.P. Swerdlow, H.R. Harke, J.Z. Zhang, N.J. Dovichi, *J. Chromatogr.*, **559** (1989), 237.
40. J.Y. Zhao, K.C. Waldron, J. Miller, J.Z. Zhang, H. Harke, N.J. Dovichi, *J. Chromatogr.*, **608** (1992), 239.
41. J.Y. Zhao, N.J. Dovichi, O. Hindsgaul, S. Gosselin, M.M. Palcic, *Glycobiology*, **4** (1994), 239-242.

CHAPTER 2

The Effect of Total Percent Polyacrylamide in Capillary Gel Electrophoresis for DNA Sequencing of Short Fragments¹

2.1 Introduction

DNA sequencing is based on the separation of labelled DNA fragments by denaturing gel electrophoresis. The rate of separation of the fragments is proportional to the electric field strength; high electric fields produce fast separations. In practice the applied field strength is limited by Joule heating which causes temperature gradients in the gel. Because the mobility of DNA fragments depends strongly on temperature, 2.3% per degree, thermal gradients lead to band broadening which degrades the separation efficiency in slab gels at high electric field strengths (1).

Conventional slab gels are about 0.5 mm thick. Thin slab gels, about 0.1 mm thick, generate fast and efficient separations at high fields (2-4). However, difficulties in automation, in maintaining a uniform gel thickness across the slab, and in detection have retarded widespread applications of this technology. On the other hand, capillary gel electrophoresis offers highly uniform chambers and high sensitivity detection technology. Typical fused-silica capillaries of 50 μm inner diameter produce outstanding thermal properties. Finally, the highly flexible nature of the capillaries simplifies automation.

Several groups have developed DNA sequencers based on capillary gel electrophoresis and laser-induced fluorescence detection (4-11). In these systems, the capillaries are filled with denaturing polyacrylamide gels. The sequencing rate observed in these gels depends on the details of the gel composition. This chapter

¹A version of this chapter has been published in *Journal of Chromatography*, 608 (1992),143-150.

describes and models the sequencing rate, resolution, and separation efficiency in DNA sequencing by capillary gel electrophoresis.

2.2 Experimental

2.2.1 Instrument Design

A schematic of the instrument is shown in Figure 2.1a. Polyimide coated fused silica capillaries (Polymicro, Phoenix, AZ, USA) were typically 35 cm long, 190 μm outer diameter (o.d.) and 50 μm inner diameter (i.d.). The capillaries were filled with gel as described below. A Plexiglas box equipped with a safety interlock holds the injection end of the capillary and the high voltage electrode. Negative high voltage is applied to the capillary to drive the DNA fragments through the capillary. The other end of the capillary is inserted into a quartz sheath flow cuvette (Precision Cells, NY, USA) (figure 2.1b). The sheath flow cuvette has 2 mm thick quartz walls and a 200 μm square flow chamber. The cuvette sits in a locally constructed holder which is held at ground potential. The sheath flow buffer, supplied by a high pressure syringe pump (Isco), flows into the cuvette around the capillary, and out the bottom. Fluorescence is excited with a 10 mW argon ion laser beam (Uniphase, CA, USA) operating at 488 nm, which passes just below the tip of the capillary (Figure 2.1b). As the DNA fragments exit the capillary they are entrained in the flowing sheath stream and pass through the laser beam. The fluorescence is collected at 90 degrees to the laser beam with a microscope objective (32X, 0.6 N.A., Leitz/Wild, Calgary, AB, Canada), imaged onto a 0.75 mm pinhole, passed through a 525 nm bandpass filter (Omega, VT, USA) to a 525 nm dichroic filter. Long wavelengths are transmitted, short wavelengths reflected. In both spectral channels the signals are collected with photomultiplier tubes (Hamamatsu, CA, USA). The current from each photomultiplier tube is dropped

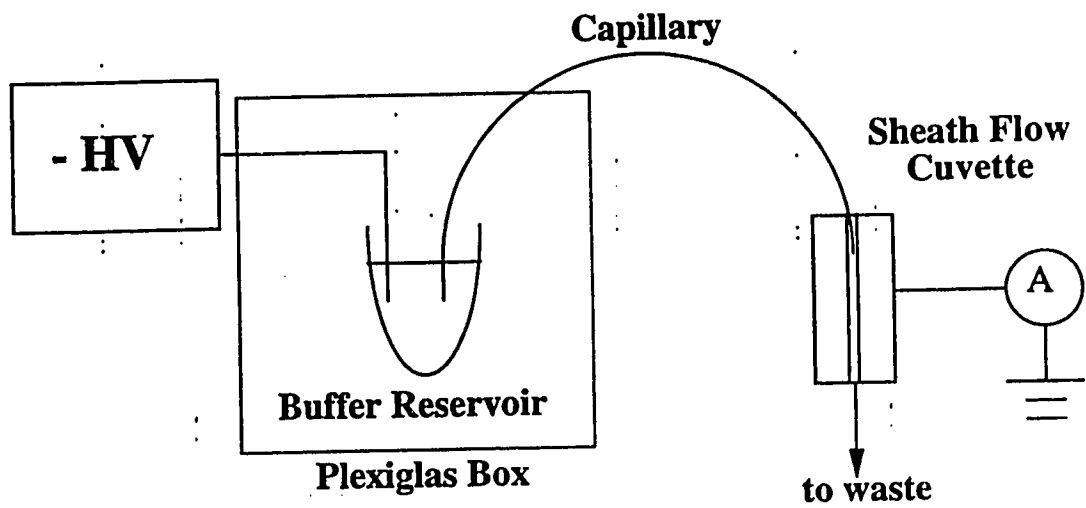


Figure 2.1a Schematic diagram of the capillary electrophoresis instrument.

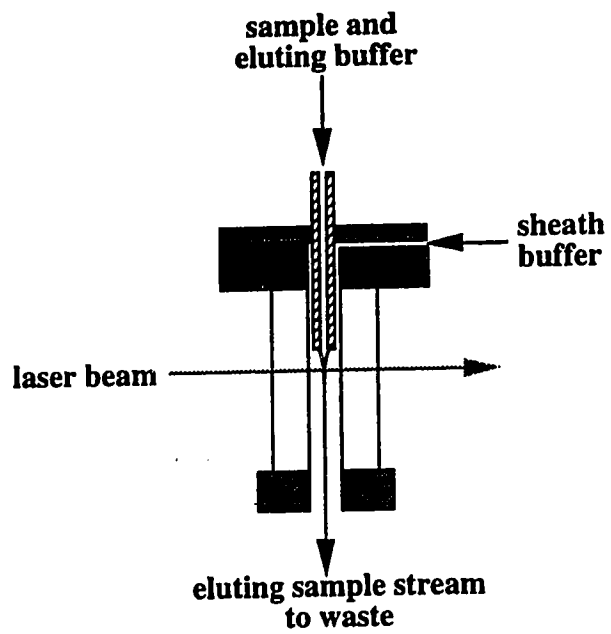


Figure 2.1b The sheath flow cuvette.

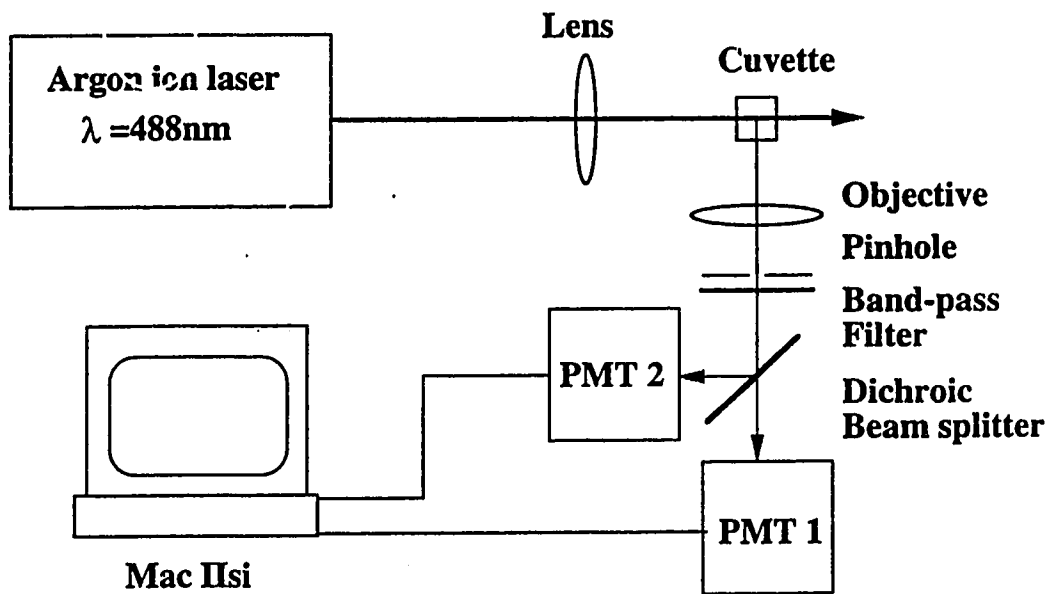


Figure 2.2 Schematic diagram of the one-laser two-channel fluorescence detector.

across a resistor, digitised, and recorded with a Macintosh IIsi computer. For this study, only the signal in the transmitted channel was used.

2.2.2 Gel Preparation

Gels were prepared in 5 ml aliquots from mixtures of acrylamide and N,N'-methylenebisacrylamide (Bis) (Bio-Rad, Toronto, ON, Canada), 1X TBE (89 mM Tris (ICN, Montreal, PQ, Canada), 89 mM borate (ICN) and 2 mM EDTA (Sigma, St. Louis, MO)) and 7 M urea (ICN). The percentage of total acrylamide in the gel, %T, was varied from 2.5 to 6%, while the percentage of acrylamide present as crosslinker (Bis), %C, was held constant at 5%. The monomer solution was carefully degassed by application of vacuum for 20 min. while the solution was stirred. Polymerisation was initiated by the addition of 2 μ l of N,N,N',N'-tetramethylethylenediamine (TEMED) and 20 μ l of 10% ammonium persulphate. The gel solution was injected into the capillary with a modified syringe. The gel was covalently bound to the last *ca.* 5 cm of the capillary with a solution of 5% (v/v) γ -methacryloxypropyltrimethoxysilane in 50% acetic acid. The end of the capillary was dipped in the silanising solution for approximately 30 sec. so that the solution filled about 5 cm of the capillary. The capillary was then filled with the gel solution. Binding the gel to the capillary walls prevented it from extruding out of the capillary into the detection cuvette. Even though polymerisation appeared to be complete within 30 minutes, gels were typically stored overnight before use. The capillaries were inspected under a microscope to ensure that they were free of bubbles before installing them in the capillary electrophoresis apparatus.

2.2.3 Sample Preparation

The DNA sequencing reaction was carried out in 40 mM Tris-HCl, pH 7.5, 20 mM MgCl₂, and 50 mM NaCl. Three microlitres of M13mp18 single stranded template DNA (1 mg/ml) was mixed with 25 ng of -21 M13 universal primer (United States Biochemical Corp., Cleveland, OH, USA). Deoxynucleotide triphosphates (dNTPs) and fluorescently labelled dideoxyadenosine triphosphate (ddATP) (DuPont, DE, USA) were added in a 10:1 ratio of dNTP:ddATP. Ten units of Sequenase 2.0 (USB) were added. The sample was incubated at 37°C for 10 min to carry out the chain extension and termination reactions, then at 90°C for 10 minutes to inactivate the Sequenase. The sample was then passed through a Sephadex G-50 spin column (Sigma, St. Louis, MO) to remove the unincorporated labelled ddATP, precipitated with 98% ethanol, and redissolved in 4.0 µl of a mixture of formamide-0.5 M EDTA (49:1), pH 8.0.

2.2.4 Electrophoresis

Samples were injected for 30 seconds at 150 V/cm. The sample was replaced with a fresh vial of 1X TBE and electrophoresis continued at 150 V/cm. The sheath flow stream was 1X TBE at a flow rate of 0.10 ml/h. Time was measured from the application of the separation potential. Two to three replicate runs were done on each capillary.

2.3. Results and Discussion

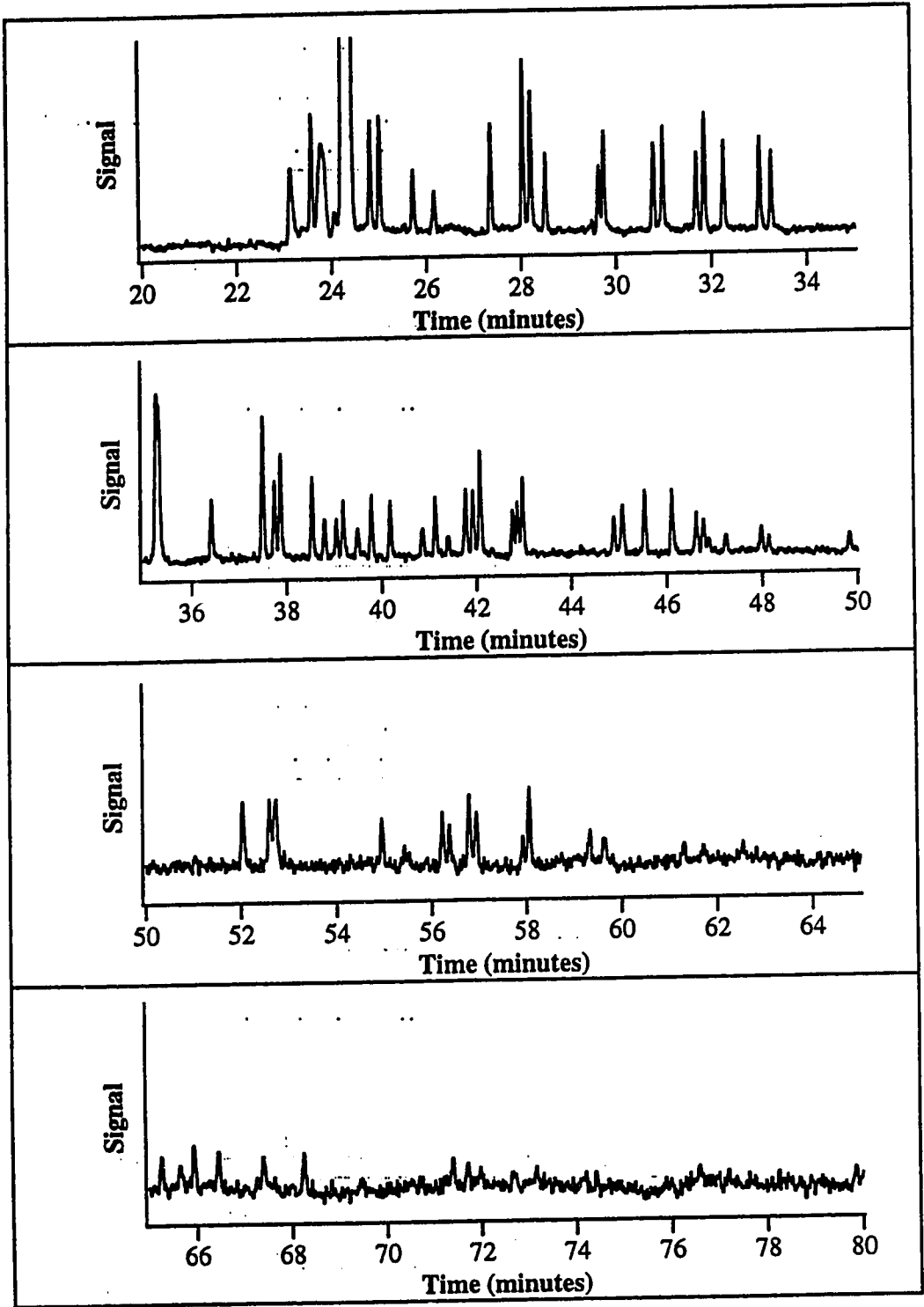
2.3.1 Performance and Stability of the Gel Filled Capillaries

Of the gel filled capillaries prepared in the manner described above, about 95% were useable. The 5% that were spoiled generally contained bubbles in the interior of the capillary. Bubbles in the gel interfered with electrophoresis since no current would flow through the capillary.

To prevent the gel from migrating out of the capillary into the detection cuvette under the influence of the high voltage, the inner wall of the capillary was treated with a silanising agent. The silanising agent, γ -methacryloxypropyltrimethoxysilane covalently bound the gel to the capillary walls. Treatment of the whole length of the capillary with the silanising agent was effective only if the total capillary length was less than 30 cm. When silanised capillaries longer than 30 cm were filled with acrylamide solution, small bubbles formed at regular intervals as the polyacrylamide gel shrank during polymerisation. Shrinkage of the gel at the ends of the shorter capillaries was presumably adequate to avoid the formation of bubbles. Silanisation of the last 2 to 5 cm of the detection end of the capillary prevented migration of the gel while allowing the preparation of capillaries longer than 30 cm.

On average, three sample injections were made on each gel filled capillary used in this study. The injection end of each capillary was trimmed between sample injections to minimize formation of bubbles at the end of the capillary. The gels were generally stable for the duration of three runs.

Figure 2.3 Electropherogram of an A-terminated M13mp18 sequencing sample.



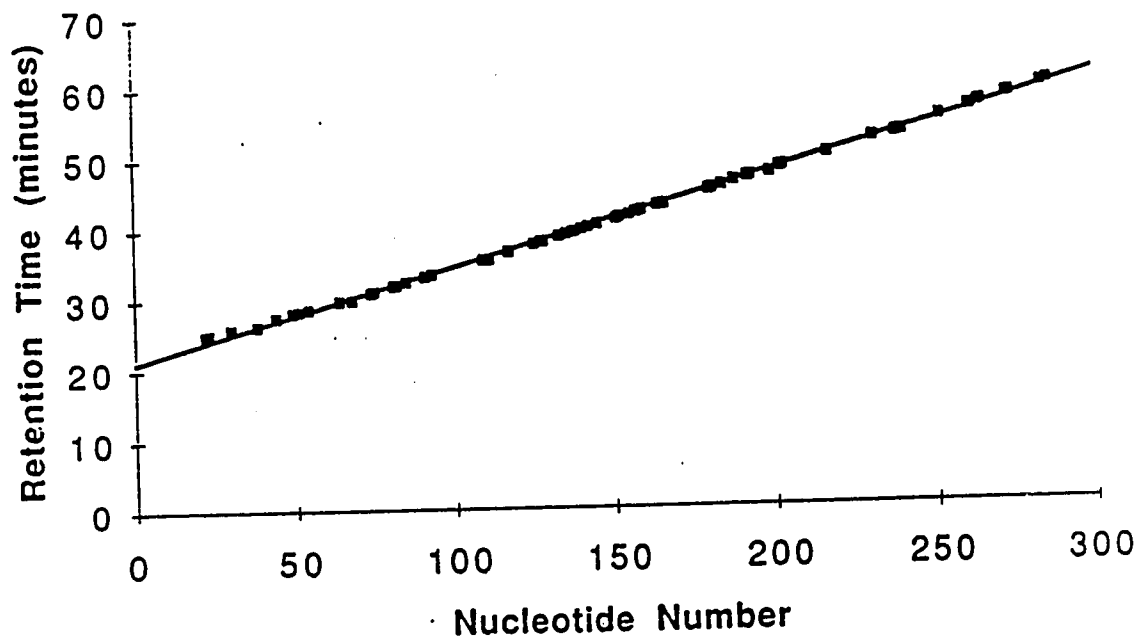


Figure 2.4 Retention time as a function of fragment length for single stranded DNA sequencing fragments separated in a 4%T, 5%C gel at an electric field strength of 150 V/cm in a 50 μm i.d., 35 cm long fused silica capillary. The separation was carried out at room temperature.

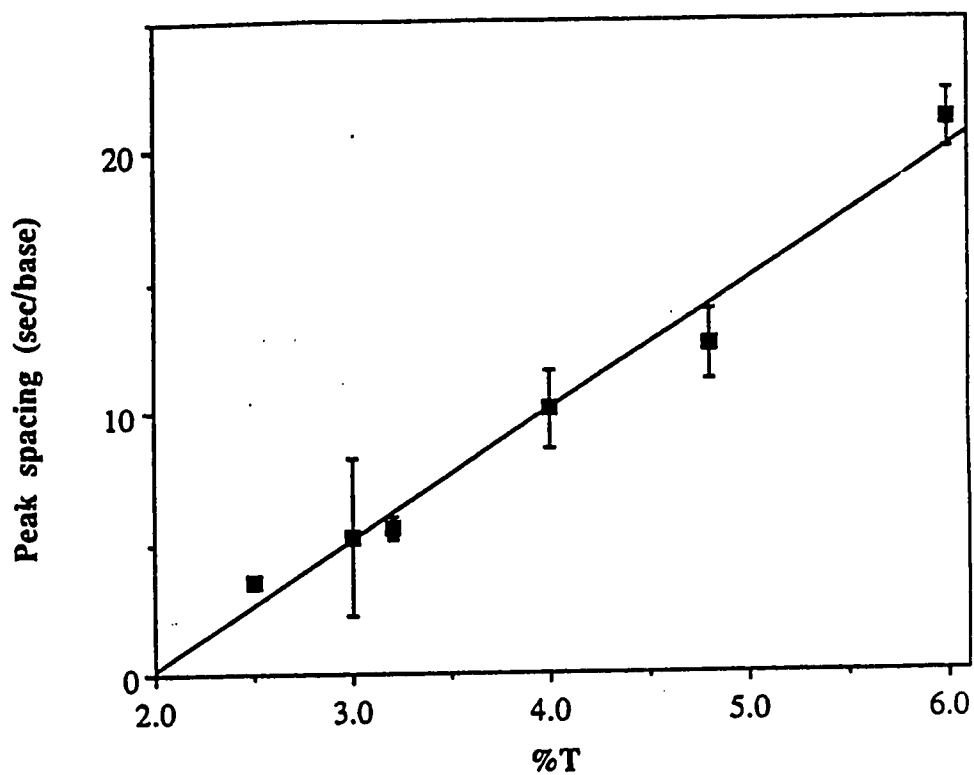


Figure 2.5 Peak spacing of DNA fragments for denaturing polyacrylamide gels with constant 5%*C*, room temperature operation, an electric field strength of 150 V/cm, and a 35 cm long 50 μm i.d. capillary. The data are shown at the 95% confidence interval and the line is the unweighted least squares fit with a linear function.

2.3.2 Sequencing Rate and Retention Time

Analysis was limited to 250 bases by the signal-to-noise ratio produced by this DNA sample. Figure 2.3 shows a typical electropherogram. The pattern of the peaks corresponds to the distribution of A's in the DNA sequence. In all cases, a plot of retention time vs. fragment length (in bases) was linear ($r > 0.998$) for fragments ranging in size from the primer to at least 250 bases; that is,

$$\text{retention time} = t_0 + M \cdot \text{spacing} \quad (2-1)$$

where t_0 is the intercept, spacing is the peak spacing in s/base, and M is the fragment length in bases. Figure 2.4 presents typical data for a 4%T gel; for these data, t_0 is 21.0 min. and the peak spacing is 8.0 s/base.

The peak spacing increased linearly with %T for all fragments (Figure 2.5). The data are shown at the 95% confidence interval; the line is the result of an unweighted least-squares fit. Only two runs were made with the 3% gel, resulting in a large confidence interval. The slope of this line, 5.0 s/base per %T, implies that the peak spacing increases by 5 s for every 1% increase in total acrylamide concentration. The intercept, -9.9 s/base, should be related to the free solution mobility of the DNA fragments. The negative sign implies that longer fragments will have a higher mobility than shorter fragments. The data of Kambara *et al.* (12) show a quadratic dependence of peak spacing on %T from 2 to 12%T. From 2 to 6%T, their data are similar to ours.

The inverse of peak spacing is sequencing rate (Figure 2.6). The data in bases/h are shown at the 95% confidence interval. The sequencing rate observed for the 2.5% gel was 1040 ± 10 bases/h, which is equal to the highest speed DNA

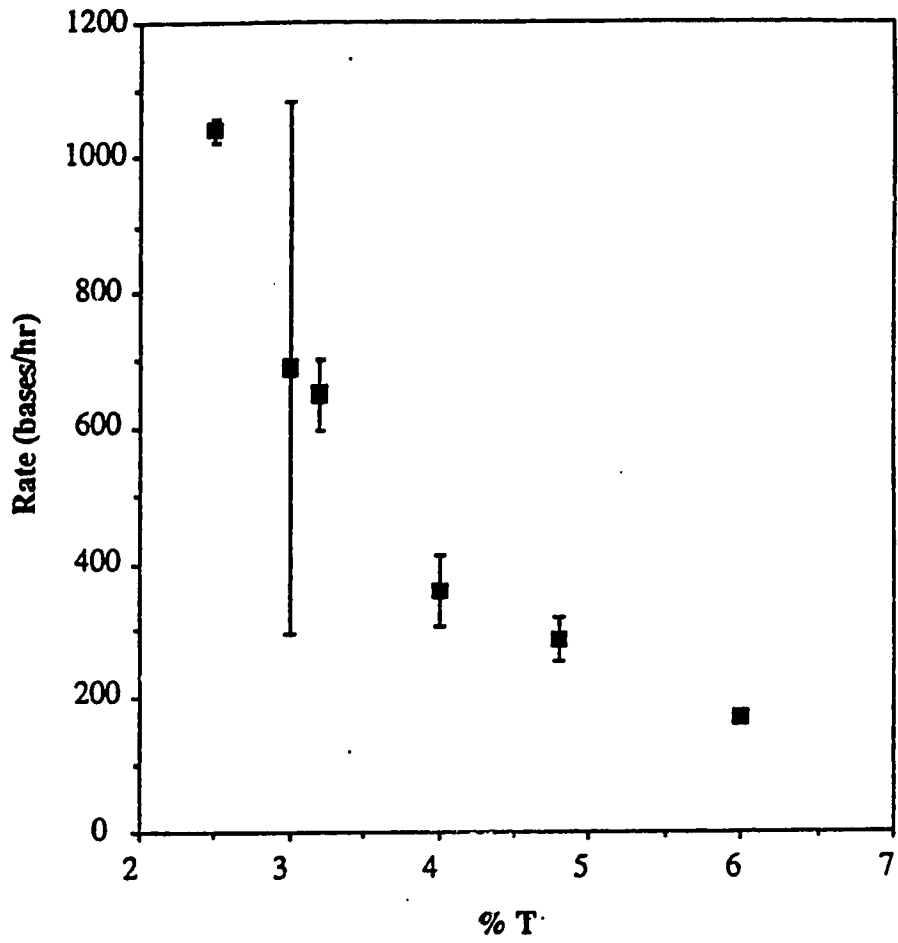


Figure 2.6 Sequencing rate for the data of Figure 2.4. Data shown at the 95% confidence interval.

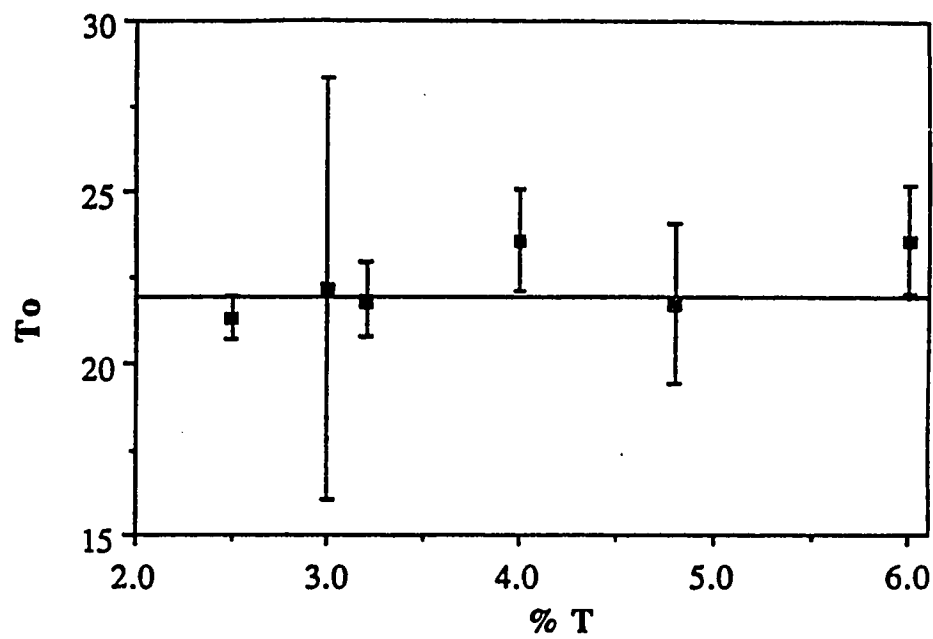


Figure 2.7 Retention time for a vanishingly small DNA fragment as a function of total acrylamide concentration. Data shown at the 95% confidence interval.

sequencing rates reported in the literature for both capillary and slab gel electrophoresis (2,11).

Stegemann *et al.* (2) reported a sequencing rate of 1000 bases/h for a slab gel separation at an electric field strength of 80 V cm^{-1} . Their data appears to have been generated at 50°C while our data were taken at room temperature, 20°C . The difference in electric field strength necessary to obtain this sequencing rate is due to the $2.3\%^\circ\text{C}^{-1}$ change in mobility with temperature (12)

The retention time for a vanishingly small DNA fragment, T_0 , is independent of %T (Figure 2.7); that is, the retention time of a hypothetical 0-base fragment does not change with %T. This independence of migration rate is not surprising because the 0-base fragment would experience no retardation by the gel. Similar data are reported by Kambara *et al.* (12) for separations of single-strand DNA performed on slab gels at an electric field of 50 V cm^{-1} . The weighted average t_0 is $21.9 \pm 0.2 \text{ min}$, shown as the horizontal line in the figure. Again, the data are shown at the 95% confidence interval. One datum was Q-tested at the 90% confidence interval from the 6% gel data set.

Combining the results from Figures 2.5 and 2.7, the retention time for a DNA fragment in a 5% C gel at 150 V cm^{-1} in a 35 cm long capillary at room temperature can be written as

$$\text{retention time} = 21.9 \text{ min} + \%T \cdot 0.083 \text{ min base}^{-1} \%T^{-1} \quad (2-2)$$

for fragments ranging from 25 to 250 bases in length, gels ranging from 2.5 to 6%T, and room temperature operation. This formula was generated from over 35 electropherograms and represents over 80 h of instrument time.

2.3.3 Electrophoretic Mobility

The relationship presented in equation 2-2 is quite robust in our laboratory. The equation is used to predict the electrophoretic mobility of a DNA sequencing fragment as a function of gel composition and nucleotide size. The electrophoretic mobility, μ , of a fragment of size M is given by

$$\mu_M = \frac{L/E}{\text{retention time}} \quad (2-3)$$

where L is the length of the capillary and E is the electric field; the ratio $L/E = 0.233 \text{ cm}^2 \text{ V}^{-1}$ for our experimental conditions. The retention time relationship of equation 2-2 is substituted into equation 2-3 to predict the electrophoretic mobility of DNA fragments ranging in size from 25 to 250 bases and for gels ranging from 2.5 to 6%T.

$$\mu_M = \frac{0.233 \text{ cm}^2 \text{ V}^{-1}}{1300\text{s} + M \cdot (-9.8\text{s base}^{-1} + 5.0\text{s base}^{-1} \%T^{-1})} \quad (2-4)$$

Equation 2-4 is a fundamental description of the electrophoretic behaviour of nucleotides in 5%C polyacrylamide gels. This behaviour must be described accurately for any successful theoretical description of DNA sequencing by gel electrophoresis.

Figure 2.8 presents the predicted mobility for DNA fragments ranging from 25 to 200 bases in length and gels ranging from 2.5 to 6%T. The data extrapolate to a mobility of $1.8 \cdot 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ for a fragment of zero bases, independent of percentage acrylamide. The mobilities that we obtain are approximately a factor of two smaller than that reported by Holmes and Stellewagen (13) for double-stranded DNA separated at an electric field strength of 3.3 V cm^{-1} and at room temperature. The results presented here are about 75% of the values reported by Kambara *et al.*

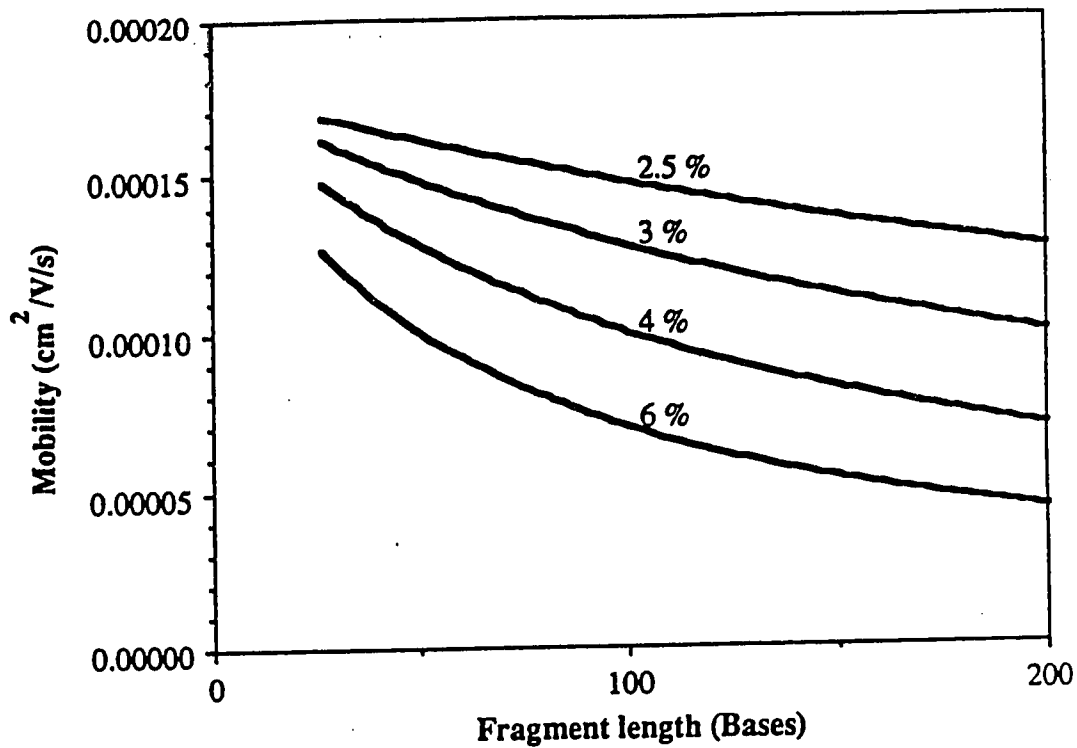


Figure 2.8 Mobility of fluorescently labelled DNA fragments. Total percent acrylamide in the sequencing gel is noted above each curve.

(12) for separation of single-stranded DNA in a slab gel at 48°C. The thermal coefficient of mobility, $2.3\%C^{-1}$, exactly accounts for the observed difference between the data of Kambara *et al.* (12) and that presented here.

According to the Ogston model (13), mobility may be written as

$$\ln \mu_M = \ln \mu_0 - K_{RM} \cdot \%T \quad (2-5)$$

Using equation 2-4 to calculate mobility, Ferguson plots (13) were generated in the range of 2.5 to 6%T for fragments ranging in size from 25 to 100 bases. The Ogston model was fitted to the data. The curves in figure 2.8 extrapolate to a common intercept, $\ln \mu_0 = -8.41 \pm 0.06$, corresponding to a mobility in free solution of $2.2 \pm 0.1 \cdot 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$. This result is 25% higher than that estimated above from the peak spacing data and one half the value reported for double-stranded DNA (13).

According to the Ogston theory, the retardation of fragments is related to the fractional volume of space in the matrix that is accessible to the analyte (13). To find the %T that produces the same pore size as a particular DNA fragment, the gel composition is found that produces a mobility one-half of that estimated for free solution. Taking the free solution mobility as $2.2 \cdot 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, the mobility of interest is $1.1 \cdot 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, or $\ln \mu = -9.12$. A 3.5%T gel is estimated to have a pore size of 3.5 nm, equal to the radius of a 100-mer single-stranded fragment; a 4.1%T gel has a pore size of 3.2 nm, equal to the radius of a 75-mer, and a 5.2%T gel has a pore size of 2.8 nm, equal to the radius of a 50-mer. The DNA fragment size is

$$\text{radius} = 0.755(\text{bases})^{1/3} \quad (2-6)$$

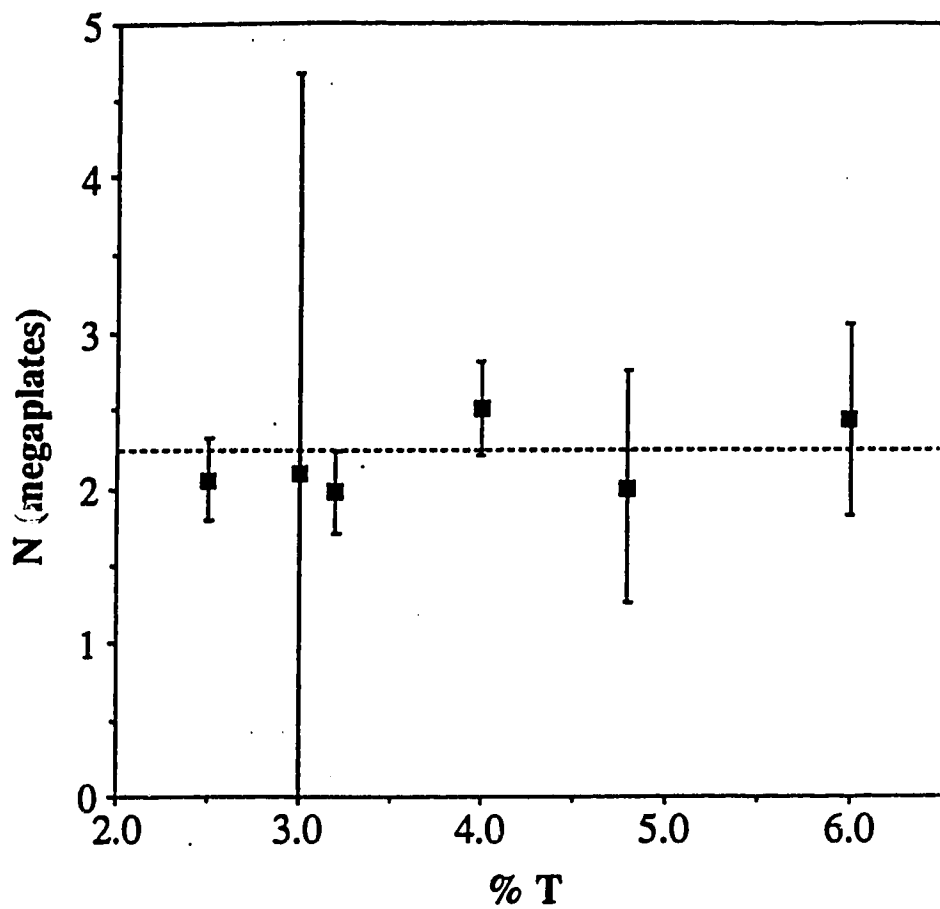


Figure 2.9 Plate count for an 85 base DNA sequencing fragment as a function of total acrylamide concentration. Data are shown at the 95% confidence interval; the dotted line is the weighted average plate count.

and is based on the geometric mean radius for double-stranded DNA (13). A plot of log pore size vs. %T is linear ($r=1.000$) over the range studied, with slope -0.56 . These data are consistent with those observed for Ferguson plots produced on slab gels with double-stranded DNA (13).

2.3.4 Theoretical Plates

The theoretical plate count for base 85 was independent of %T over the range studied, with a weighted average of $2.2 \pm 0.2 \cdot 10^6$ plates (Figure 2.9). The data are shown at the 95% confidence interval and the dashed line is the weighted average plate count. Fragments 231 bases in length had a separation efficiency, $N=2.2 \pm 0.4 \cdot 10^6$ plates, that also was independent of %T. Plate counts were estimated from the second moment calculation. Slab gel data shows a similar independence of plate count from gel composition (12), although plate count appears to increase slightly with fragment length.

2.3.5 Band Broadening - injection, detection, and thermal gradient

It is interesting to speculate on the origin of the constant plate count. The product of injection voltage and time was varied by several orders of magnitude to determine the effect of column overloading; no improvement in plate count was noted for the smallest sample loadings. Detection time constant and volume do not seem to be important; the system response time would limit plate counts to *ca.* 100 million.

Thermal band broadening could contribute to plate count. Joule heating will produce a parabolic temperature profile in the capillary. For all the gels studied, the

electric current was constant, 1.61 ± 0.01 mA, and independent of gel composition.

The heat generated per unit volume in a capillary of radius r is

$$Q = \frac{EI}{r^2} = \frac{\lambda CV^2}{L^2} = \frac{150 \text{ V cm}^{-1} \cdot 1.6 \cdot 10^{-6} \text{ A}}{\pi (2.5 \cdot 10^{-3} \text{ cm})^2} = 1.22 \cdot 10^1 \text{ W cm}^{-3} \quad (2-7)$$

Note that the heat dissipated in the capillary scales with voltage squared, at constant molar conductance, λ , and ionic strength, C . Knox (14) has stated that the temperature difference between the axis of the capillary and the inner wall, θ , is given by

$$\theta = \frac{Qr^2}{4\kappa} = \frac{1.22 \cdot 10^1 \text{ W cm}^{-3} \times (2.5 \cdot 10^{-3} \text{ cm})^2}{4 \times (4 \cdot 10^3 \text{ W cm}^{-1} \text{ K}^{-1})} = 0.005 \text{ K} \quad (2-8)$$

where κ is the thermal conductivity of the solution, rather arbitrarily estimated as $4 \text{ mW cm}^{-1} \text{ K}^{-1}$ for the 7 M urea, 1X TBE, polyacrylamide solution. The temperature difference across the capillary is very small for $50 \mu\text{M}$ i.d. capillaries at 200 V cm^{-1} electric fields.

The parabolic temperature profile is translated to a velocity profile by the relative thermal coefficient of mobility, $(d\mu/dT)/\mu = 0.023 \text{ K}^{-1}$ (12) (where T = temperature). For the gels used in this study, the relative mobility of a fragment at the centre of the capillary will be 0.00012 (0.012%) higher than the mobility of a fragment at the capillary wall. The maximum theoretical plate count in the presence of this thermally induced velocity profile is related to the diffusion coefficient of the DNA fragment. From the Stokes-Einstein formula, the diffusion coefficient of a molecule is given by (15)

$$D_m = \frac{kT}{6\pi\eta r_{\text{molecular}}} \quad (2-9)$$

where k is the Boltzmann constant. The data of Nishikawa and Kambara (1) suggest that the diffusion coefficient of a 100-mer fragment is about $1.0 \cdot 10^{-10} \text{ cm}^2 \text{ s}^{-1}$ at 50°C . Converting to 20°C , the diffusion coefficient is expected to be about $9 \cdot 10^{-8} \text{ cm}^2 \text{ s}^{-1}$. Substituting the diffusion coefficient, capillary length, fraction velocity difference, capillary radius, electrophoretic mobility, and electric field strength into Knox's equation

$$\begin{aligned}
 N_{\text{thermal}} &= \frac{24D_m L}{\left(\frac{d\mu/dT}{\mu} \theta\right)^2 r^2 \mu E} \\
 &= \frac{24 \cdot 9 \cdot 10^{-8} \text{ cm}^2 \text{ s}^{-1} \cdot 35 \text{ cm}}{\left[(0.00012)^2 (2.5 \cdot 10^{-3} \text{ cm})^2 (1.0 \cdot 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \cdot 150 \text{ V cm}^{-1})\right]} \quad (2-10) \\
 &= 5.6 \cdot 10^{10}
 \end{aligned}$$

yields $N_{\text{thermal}} = 56$ billion theoretical plates for an 86-mer fragment. Thermally induced band broadening does not appear to be significant in this capillary system.

N_{thermal} scales inversely with the fifth power of electric field and thermally induced band broadening is insignificant except at very high electric fields in capillaries. For example, an electric field strength of 800 V cm^{-1} will produce a temperature difference of 0.2°C across the $50 \mu\text{m}$ i.d. capillary, corresponding to $N_{\text{thermal}} = 5 \cdot 10^6$ plates. On the other hand, the thermal plate count scales inversely with the sixth power of radius. A 0.5 mm diameter capillary (with similar thermal characteristics as a 0.5 mm thick slab) will have a limiting plate count of 56 000 when operated at 150 V cm^{-1} . Thermal gradients dominate the performance of conventional slab gels at high electric fields.

2.3.6 Band Broadening - longitudinal diffusion

Longitudinal diffusion appears to dominate band broadening in gel filled capillaries. Using the electrophoretic mobility of 85-mer fragments in 4%T gels, the plate count due to longitudinal diffusion is given by

$$N_{\text{longitudinal}} = \frac{\mu V}{2D_m} = \frac{1.0 \cdot 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \cdot 5250 \text{ V}}{2 \cdot 9 \cdot 10^{-8} \text{ cm}^2 \text{ s}^{-1}} = 2.9 \cdot 10^6 \quad (2-11)$$

which, given the assumptions in the estimation of diffusion coefficient, is in remarkable agreement with the data. The dependence of plate count on fragment size can be estimated from the mobility formula of equation 2-4, the size dependence of DNA from equation 2-6, and the longitudinal diffusion equation above, and can be written as

$$\begin{aligned} N_{\text{longitudinal}} &= \frac{\mu V}{2D_m} = \frac{0.233V}{1300s + M \cdot (-9.8s \text{ base}^{-1} + 5.0s \text{ base}^{-1} \cdot \%T)} \\ &\quad \frac{2kT}{6\pi\eta 0.755M^{1/3}} \\ &= \frac{AM^{1/3}V}{1300s + M \cdot (-9.8s \text{ base}^{-1} + 5.0s \text{ base}^{-1} \cdot \%T)} \end{aligned} \quad (2-12)$$

where A is a constant related to temperature and cross-linker concentration. For an 85-mer in a 4%T gel at a voltage of 5250 V, $A \approx 2.7 \cdot 10^5 \text{ s V}^{-1}$. A plot of expected plate count vs. fragment length is shown in Figure 2.10. Plate count maximizes for short fragments, ca. 85-mer, but varies by only 25% for fragments ranging in size from 25 to 250 bases. Within experimental error, plate count is independent of fragment length.

Plate count should increase linearly with applied potential. Potential can be increased either by increasing the electric field strength, for a constant length capillary,

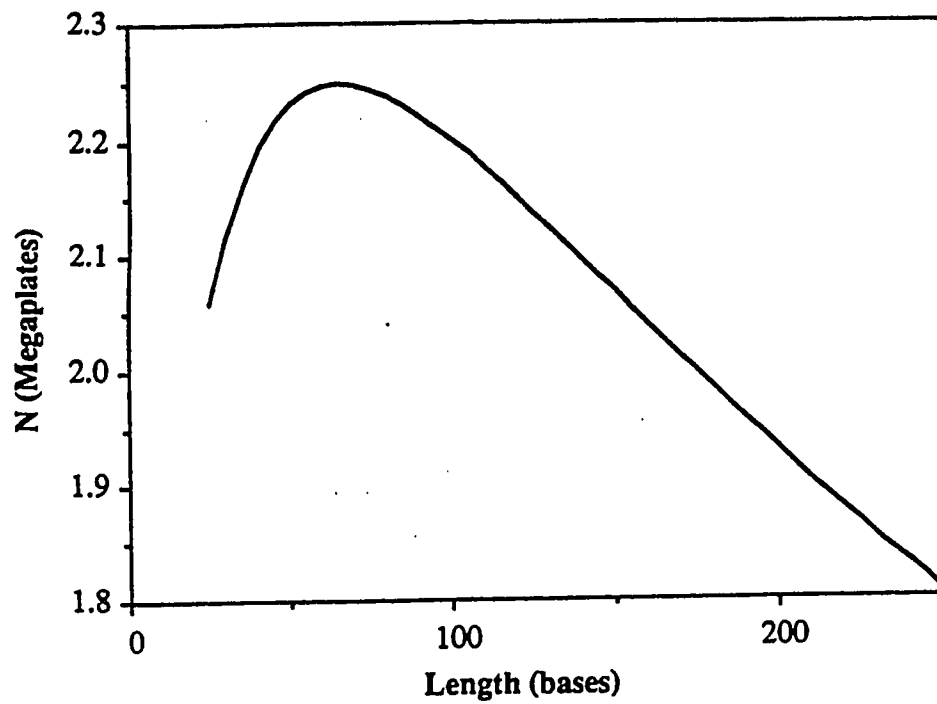


Figure 2.10 Predicted plate count in a 4% total acrylamide gel.

or by increasing the length of the capillary, at constant electric field. The electric field strength cannot be increased without bounds. Polyacrylamide gels are unstable at electric field strengths greater than about 500 V cm^{-1} . Large bubbles form in the capillary at high fields, destroying the separation. However, microbubbles may form at intermediate potentials, creating eddy diffusion and degrading the separation efficiency. It appears that longer capillaries are required to produce very high separation efficiency, albeit at the expense of longer analysis time.

2.3.7 Resolution

Resolution was determined graphically for peaks 85-86 at different %T. The data are shown at the 95% confidence interval in Figure 2.11. The range of resolution observed for the data, from 1 to 2, is quite similar to results from slab gel data (12). Resolution degraded with increased length of sequencing fragment, for all %T studied; a similar phenomenon is present in the data of Kambara *et al.* (12) for fragments ranging in size from 100 to 400 bases.

Resolution of adjacent peaks is related to the theoretical plate count and the relative peak spacing (16,17)

$$\text{resolution} = \frac{\sqrt{N}}{4} \cdot \frac{\text{spacing}}{t_r} \quad (2-13)$$

Substituting expressions for plate count, peak spacing, and retention time, the resolution is predicted to be

$$\text{Resolution} = \frac{\sqrt{A \times V \times M^{1/3}} \times \frac{1}{4} \times (-9.84 \text{sec base}^{-1} + 4.99 \text{sec base}^{-1} \times \%T)}{\left\{1310 \text{sec} + M \times (-9.84 \text{sec base}^{-1} + 4.99 \text{sec base}^{-1} \times \%T)\right\}^{3/2}} \quad (2-14)$$

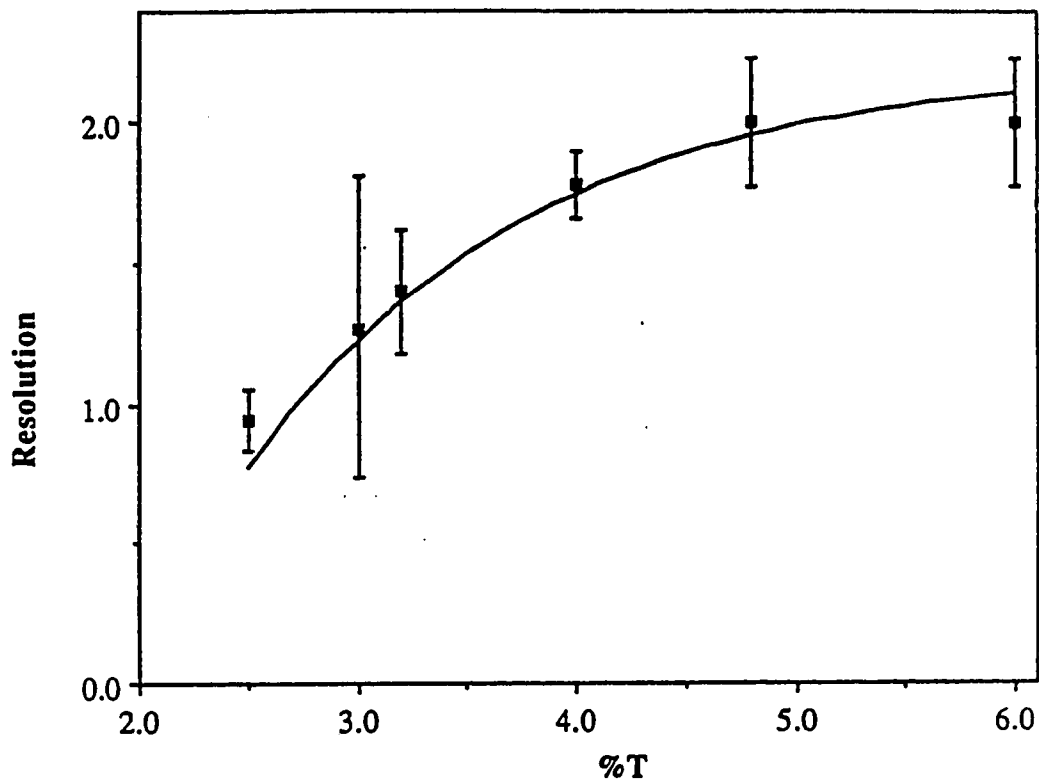


Figure 2.11 Resolution of bases 85-86. Data are shown at the 95% confidence interval. The smooth curve is the prediction of equation 2-14; no parameters were adjusted.

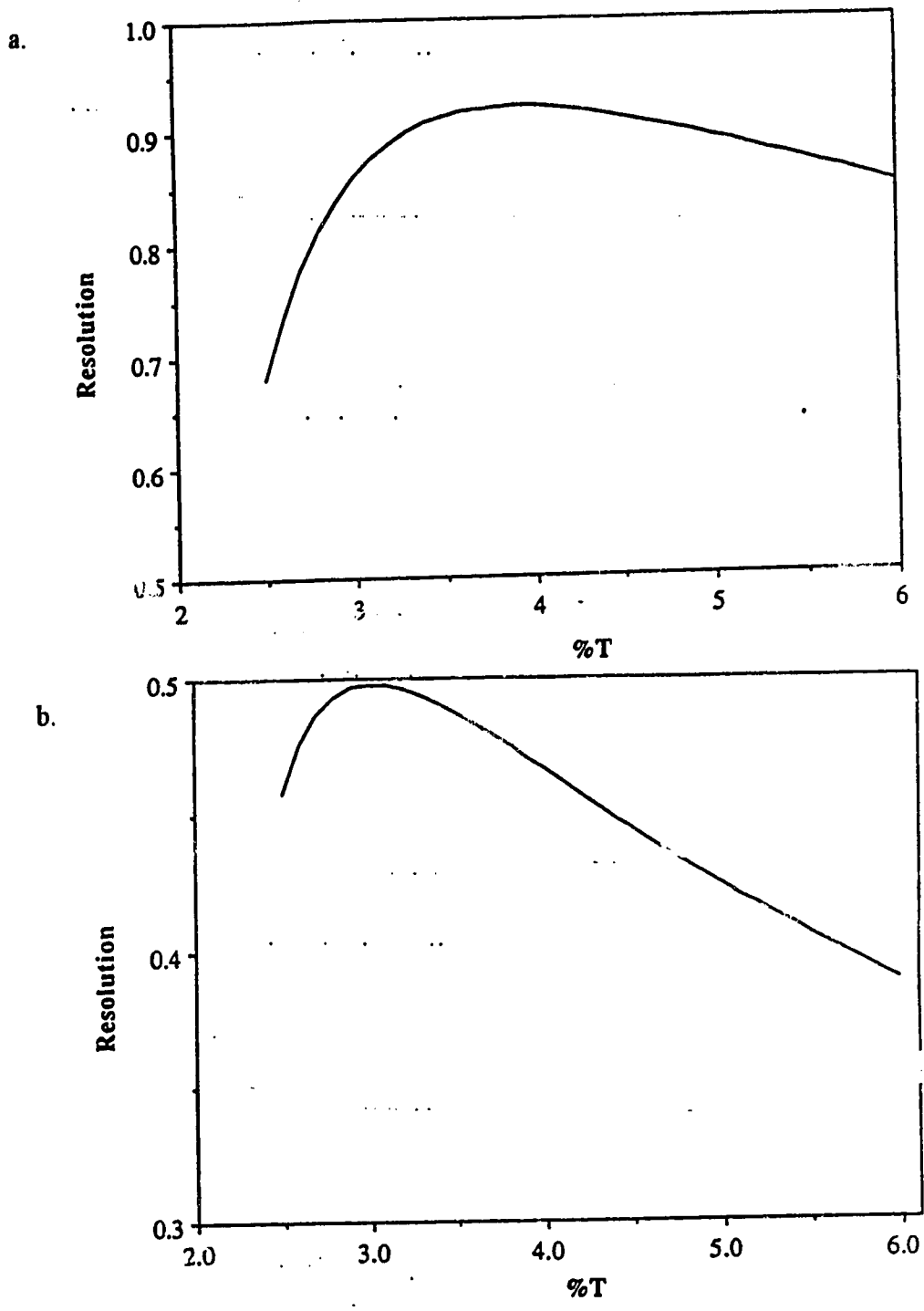


Figure 2.12 Predicted resolution for adjacent DNA fragments as a function of total acrylamide concentration. a. Fragments 250-251 bases in length. b. Fragments 500-501 bases in length.

The smooth curve in Figure 2.11 is a plot of predicted resolution for a fragment 82 bases in length and a potential of 5250 V. Recalling that there are no adjustable parameters in the theory, the agreement with the data is outstanding.

Resolution decreases as *ca.* $M^{-3/2}$ for larger fragments at constant gel composition. For any given length fragment, resolution is maximized for a particular gel composition. Figure 2.12 presents a plot of resolution vs. %T for a fragment 250 bases in length. Under the conditions at which the gels are run, the optimum resolution for a 250-mer is produced with a 4%T gel. To analyse long DNA fragments, it is appropriate to use low %T gels.

2.4 Conclusions

This chapter presents phenomenological models for DNA sequencing in polyacrylamide gels with 5% cross-linker concentration. The mobility model has three parameters: fragment length, %T, and the retention time of a vanishingly small DNA fragment. The plate count and resolution model contains one additional parameter, the diffusion coefficient of a DNA fragment of known size in a gel of known concentration.

This model is limited to 5%C polyacrylamide gels ranging from 2.5 to 6%T and operating at 150V cm^{-1} at room temperature. Based on slab-gel data, equation 2-1 requires a quadratic term to describe sequencing data for gel compositions extending to 10%T. Because mobility increases by 2.3% per degree temperature rise, faster separations are expected at higher temperatures. Similarly, different cross-linker concentration or composition will lead to different sequencing rate.

The phenomenological model presented in this chapter is limited to short fragments at an electric field strength of 150V cm^{-1} . The model is expected to fail for

longer fragments and higher electric fields. Theory for double stranded DNA states that short fragments at low electric field exist in a random coil configuration (18); the mobility of these fragments scales inversely with size. At higher electric fields, or for longer fragments, there is a transition to a stretched or linear configuration; the migration of these fragments is independent of size. Of course, it is not possible to obtain sequence information from fragments in the linear configuration. For double stranded DNA, the transition from random coil to stretched rod configuration is predicted to scale as M/E^2 (18). Extrapolation of the model to fragments longer than 250 bases or for electric fields higher than 150 V cm^{-1} is not warranted.

..

References

1. Nishikawa, T., Kambara, H., *Electrophoresis*, **12** (1991), 623.
2. Stegemann, J., Schwager, C., Erfle, H., Hewitt, N., Voss, H., Zimmermann, J., Ansorge, W., *Nucleic Acids Res.*, **19** (1991), 675.
3. Kostichka, A.J., Marchbanks, M.K., Brumley, R.L., Drossman, H., Smith, L.M., *Bio/Technology*, **10** (1992), 78.
4. Swerdlow, H., Gesteland, R., *Nucleic Acids Res.*, **18** (1990), 1415.
5. Drossman, H., Luckey, J.A., Kosichka, A.J., D'Cunha, J., Smith, L.M., *Anal. Chem.*, **62** (1990), 900.
6. Cohen, A.S., Najarian, D.R., Karger, B.L., *J. Chromatogr.*, **516** (1990), 49.
7. Swerdlow, H., Wu, S., Harke, H., Dovichi, N.J., *J. Chromatogr.*, **516** (1990), 61.
8. Luckey, J.A., Drossman, H., Kostichka, A.J., Mead, D.A., D'Cunha, J., Norris, T.B., Smith, L.M., *Nucleic Acids Res.*, **18** (1990), 4417.
9. Chen, D.Y., Swerdlow, H.P., Harke, H.R., Zhang, J.Z., Dovichi, N.J., *J. Chromatogr.*, **557** (1991), 237.
10. Karger, A.E., Harris, J.M., Gesteland, R.M., *Nucleic Acids Res.*, **19** (1991), 4955.
11. Swerdlow, H., Zhang, J.Z., Chen, D.Y., Harke, H.R., Grey, R., Wu, S., Dovichi, N.J., Fuller, C., *Anal. Chem.*, **62** (1991), 2835.
12. Kambara, H., Nishikawa, T., Katayama, Y., Yamaguchi, T., *Bio/Technology*, **6** (1988), 816.
13. Holmes, D.L., Stellewagen, N.C., *Electrophoresis*, **12** (1991), 253.
14. Knox, J.H., *Chromatographia*, **26** (1989), 329.

15. Mosher, R.A., Dewey, D., Thormann, W., Saville, D.A., Bier, M., *Anal. Chem.*, **61** (1989), 362.
16. Jorgenson, J.W., Lukacs, K.D., *Anal. Chem.*, **53** (1981), 1298.
17. Giddings, J.C., *Sep. Sci.*, **4** (1969), 181.
18. Noolandi, J., *Can. J. Phys.*, **68** (1990), 1055.

CHAPTER 3

Two-Label Peak-Height Encoded DNA Sequencing by Capillary Gel Electrophoresis²

3.1 Introduction

In 1989 Tabor and Richardson reported the effect of manganese on the incorporation of dideoxynucleoside triphosphates (ddNTPs) by T7 DNA polymerase (1). Manganese increases the incorporation rate of the ddNTPs and also produces uniform termination of the DNA sequencing reactions. In 1990, both Ansorge *et al.* and Tabor and Richardson independently reported DNA sequencing protocols based on T7 polymerase with manganese (2-3). Both protocols require a single sequencing reaction with one fluorescently labelled primer. Adjustment of the ddNTP concentrations produces peak heights in a ratio of 8:4:2:1 that encode the DNA sequence. Ansorge also reported a variation of this method in which two sequencing reactions are carried out, each with the same fluorescently labelled primer: ddCTP and ddTTP are present in the first in a ratio of 2:1 while ddATP and ddGTP are present in the second again in a 2:1 ratio. The products of these reactions are separated on adjacent lanes of a polyacrylamide slab gel.

This single reaction technique offers advantages for primer walking applications. Each sample requires only one labelled primer and one reaction. In addition, the separation of the reaction products is carried out in a single lane; this increases sample throughput. Because only a single lane is required the technique is also easily adapted to capillary gel electrophoresis, which provides more rapid separation of the sequencing fragments (4-12).

²A version of this chapter has been published in *Nucleic Acids Research*, 20 (1992), 4873-4880.

The 8:4:2:1 peak height ratio leads to poor accuracy, typically 90% for fragments shorter than 250 bases (11). There are three sources of error in sequence determination by this technique. First, for each ddNTP the relative variation in peak heights can be about 25% (2). This distribution in peak amplitude leads to errors in identification of the terminal nucleotide, particularly if the peak height ratio is not carefully adjusted. Second, the smallest peaks are frequently lost when sandwiched between peaks of larger amplitude. This problem is more severe for longer fragments where the electrophoretic resolution degrades. For this reason it is difficult to sequence past 250 bases with the four peak height sequencing technique. Third, systematic errors occur due to ghost peaks that are associated with false priming, finite processivity of the polymerase, and contaminant oligonucleotides present in the sample. The ghost peaks, superimposed on the sample peaks, cause changes in the peak heights and errors in identifying the sequence. This effect is most pronounced for the lower amplitude peaks, whose height may be increased by 50 to 100% by the presence of a ghost peak. False priming and artifacts from contaminant oligonucleotides require that highly purified DNA be used for sequence determination.

This chapter describes a modification to this technique. The variation on this method described by Ansorge *et al.* (2) uses two reactions, each with two ddNTPs added in a ratio of 2:1. While it is difficult to control a peak height ratio of 8:4:2:1, a ratio of 2:1 is simple to maintain for any two ddNTPs. Two sequencing reactions are performed, each with a different fluorescently labelled primer and two ddNTPs. The reaction products are pooled and separated on a single gel filled capillary. This method combines the convenience of a single-column separation with the accuracy inherent in the two level discrimination, and the accuracy of multiple labels, at the expense of performing two sequencing reactions.

3.2 Experimental

3.2.1 Electrophoresis

The capillary electrophoresis system is described in detail in section 2.2.1. The capillaries are 50 μm i.d., 190 μm o.d., typically 35 to 40 cm long. The gels are prepared in 5 ml aliquots from carefully degassed solutions of acrylamide and bisacrylamide (4%T, 5%C), 1X TBE and 7 M urea. Polymerisation is initiated by addition of 2 ml of TEMED and 20 ml of 10% (w/v) ammonium persulfate. The gel solution is injected into the capillary with a modified syringe. The gel is covalently bound to the last 2 cm of the capillary with γ -methacryloxypropyltrimethoxysilane to prevent deformation of the gel into the detection cuvette. Polymerisation appears complete in 30 minutes, but capillaries are typically stored overnight before use.

3.2.2 Detector

The detector was initially constructed for the DuPont Genesis sequencing system (Figure 3.1) (11). Fluorescence is excited by a 30 mW argon ion laser operating at 488 nm, and is collected at 90 degrees by a 0.60 NA, 32X microscope objective (Leitz/Wild model 2569-1130). The fluorescence is imaged onto a 0.75 mm diameter pinhole. A dichroic filter splits the fluorescence into two spectral channels. The fluorescence transmitted by the dichroic filter passes through a bandpass interference filter with a 35 nm bandwidth centered at 550 nm, and detected with an R1477 photomultiplier tube (PMT). The fluorescence reflected from the dichroic filter passes through a second bandpass interference filter, centered at 515 nm and with a 25 nm bandwidth, and is detected with an R1477 PMT. The two bandpass filters

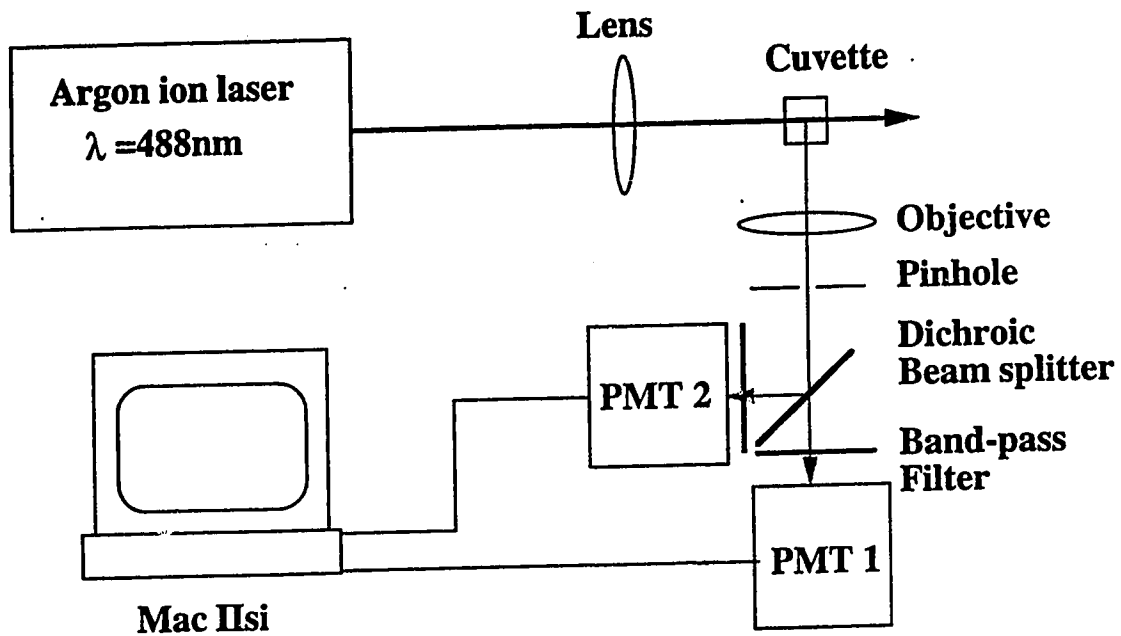


Figure 3.1 Diagram of the two-channel fluorescence detector.

replace the single bandpass filter in the original detector design (Sec. 2.2.1). The output from each PMT is conditioned with a simple low-pass electronic filter with a 0.5 sec time constant and digitised by a National Instruments A/D board in a Macintosh IIsx computer. The data are treated with a binomial filter before presentation. The sheath flow is provided by a syringe pump operating at a flow rate of 0.08 ml/h.

3.2.3 Sample Preparation

Three micrograms of template DNA was added to 2.4 pmol of FAM labelled primer (ABI -21 M13 primer), 2.5 μ l 10X MOPS buffer (400 mM MOPS, pH 7.5, 500 mM NaCl, 100 mM MgCl₂), 2.5 μ l 10X Mn solution (50 mM MnCl₂, 150 μ M sodium isocitrate) and water to a volume of 18 μ l. The primer and template were annealed by heating the mixture to 65 C for 2 minutes, followed by slow cooling to room temperature. Four microlitres of the combined T/G termination mix (the T and G termination mixes were mixed in either a 2:1 or a 3:1 ratio T:G; each mix was 1 mM each dATP, dCTP, dGTP, dTTP, and 3.3 mM of the appropriate ddNTP) was added. The mixture was warmed to 37 C for 2 minutes, then 6 units of Sequenase Version 2.0 and 0.006 units of pyrophosphatase were added and the mixture was incubated at 37 C for 10 minutes. The reaction was stopped by addition of 12 μ l stop/salt solution (20 mM EDTA, 1 M sodium acetate, pH 8.0) and the DNA was precipitated with 120 μ l 98% ethanol. The sample was kept at -20 C for at least 30 minutes, washed with 300 μ l ice cold 80% ethanol and dried under vacuum. Identical experimental conditions were used with a JOE labelled primer to yield a nominal peak height ratio of either 2:1 or 3:1 for A and C. The samples were resuspended in 3 ml of a 49:1 mixture of formamide-0.5 M EDTA and the two samples were mixed. The combined

sample was heated to 95 C for two minutes, and injected onto the capillary by applying a 200 V/cm electric field for 20 to 60 seconds.

3.2.4 Sequence Determination

The sequence was interpreted by eye from the smoothed electrophoresis data. The data were plotted in 20 minute intervals with approximately 60 peaks per plot. Two lines were drawn across the plot corresponding to the discrimination level for each fluorescent dye. For fragments between 350 and 400 bases in length, small amplitude fragments occasionally suffer from overlap with neighbouring peaks. All electropherograms were smoothed with a 2 pass binomial filter before presentation.

3.3 Results and discussion

The instrument described here was originally designed for use with the DuPont fluorescently labelled dideoxynucleosides(11). The DuPont system, which is no longer manufactured, excited fluorescence with a single argon ion laser beam and in our version, a single microscope objective collected fluorescence. A simple spectrophotometer was constructed from a dichroic filter that reflects light of wavelengths shorter than 530 nm and transmits longer wavelengths. Spectral bandpass filters isolate fluorescence in the 515 nm and 550 nm regions. Two fluorescently labelled primers (FAM and JOE from Applied Biosystems) are available that match the spectral filters used in the sequencer. The use of two separate bandpass filters provides better spectral resolution than the single bandpass filter used in the earlier version of the instrument (sec. 2.2.1). There is still some optical cross-talk; the FAM labelled primer gives signal in both channels.

The excitation laser was operated at a power of 30 mW. It was determined that the FAM labelled primer was photobleached at laser powers greater than 10 mW. There was however no reduction in signal-to-noise ratio for a laser power of 30 mW so the quality of the FAM signal was not compromised at the higher power. The maximum signal and signal to noise ratio for the JOE labelled primer was obtained at 30 mW laser power. Since JOE generally gave less signal than FAM, and since the length of sequence data, as well as the accuracy, were dependent on signal to noise ratio late in the run, 30 mW was chosen as the optimum laser power for sequencing experiments with this primer pair.

The FAM and JOE labelled primers have been used in the modified peak amplitude sequencing technique. Figures 3.2 and 3.3 show preliminary results for MMTV samples. Figure 3.2a shows the electropherogram of the FAM labelled A and C terminations; FAM gives signal in both channels but only the transmitted channel is plotted. The large, wide peak at about 30 minutes is the excess labelled primer. The DNA sequencing peaks which follow the primer decrease exponentially in intensity as the fragments become longer. The individual peaks that form the exponential envelope have uniform intensities due to the presence of manganese in the sequencing buffer (1). There are two sets of peaks in figure 3.2a. The peaks of highest intensity correspond to the A terminated DNA fragments. A second set of less intense peaks, also with uniformly decreasing intensities, are due to the C terminated DNA fragments. The peak height ratio is 2:1 for A:C.

An expansion of the electropherogram is shown in figure 3.2b. The pattern of peaks was correlated with the known sequence for MMTV for fragments 5 to 320 nucleotides longer than the primer. Some background peaks are present in this sample, for example in the region just before the A peak at 40 minutes. These ghost peaks do not interfere with the identification of the sequence as long as they are less

than 25% the intensity of the smaller peaks. The top panel of figure 3.2c shows fragments 40 to 50 nucleotides longer than the primer. The resolution of the two A peaks 41 and 42 centred at 42.5 minutes is 1.67. The bottom panel shows fragments 200 to 250. The resolution of the A peaks 200 and 201 is 0.90. There are no ambiguities in the peak height encoding. However, identification of the small peaks which code for C terminated fragments becomes difficult as the signal-to-noise ratio decreases late in the run.

Figure 3.3 shows the JOE labelled T₂₀₁ terminations. The peak height ratio is 2:1 T:G; the large peaks are T's and the small peaks are G's. The pattern of peaks was correlated with the known sequence from 5 to 200 nucleotides. Figure 3.3c shows an expansion of the regions of the electropherograms for fragments 20 to 30 and 110 to 120 nucleotides in length. The difference in peak heights between T 115 and G 116 is less than the theoretical 2:1. This would make discrimination between T and G in an unknown sample difficult. Also present in this panel are two ghost peaks of significant intensity at positions 117 and 119. These could be mistaken for G's in this electropherogram. The known sequence for MMTV shows that there are C's in positions 117 and 119.

A sequencing electropherogram of MMTV DNA is presented in figure 3.4; the peak height ratio is 2:1. Peaks in both traces are A and C; the large peaks are A, the small peaks are C. Peaks in the solid trace only are T and G; large peaks are T and small peaks are G. The sequence was called to 330 nucleotides. The agreement of the called sequence with the known sequence of MMTV was 98%. Four of the seven errors were misidentification of a T as a G due to ambiguous peak heights. Two were misidentification of a C as a G due to poor signal to noise ratio in the reflected channel obscuring the low intensity C peaks.

Figure 3.2a Electropherogram of an A and C terminated MMTV sample. The large peaks are A's, the small peaks are C's.

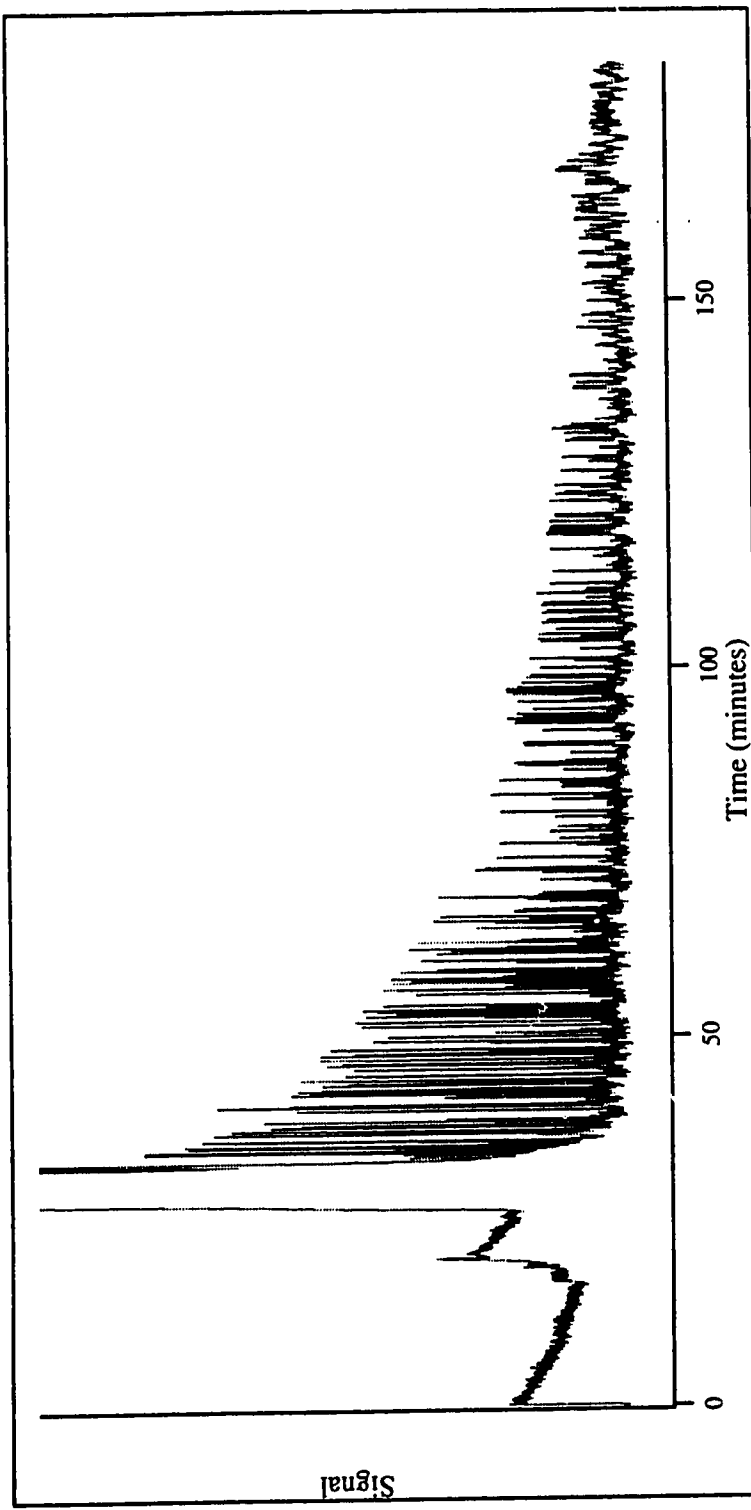
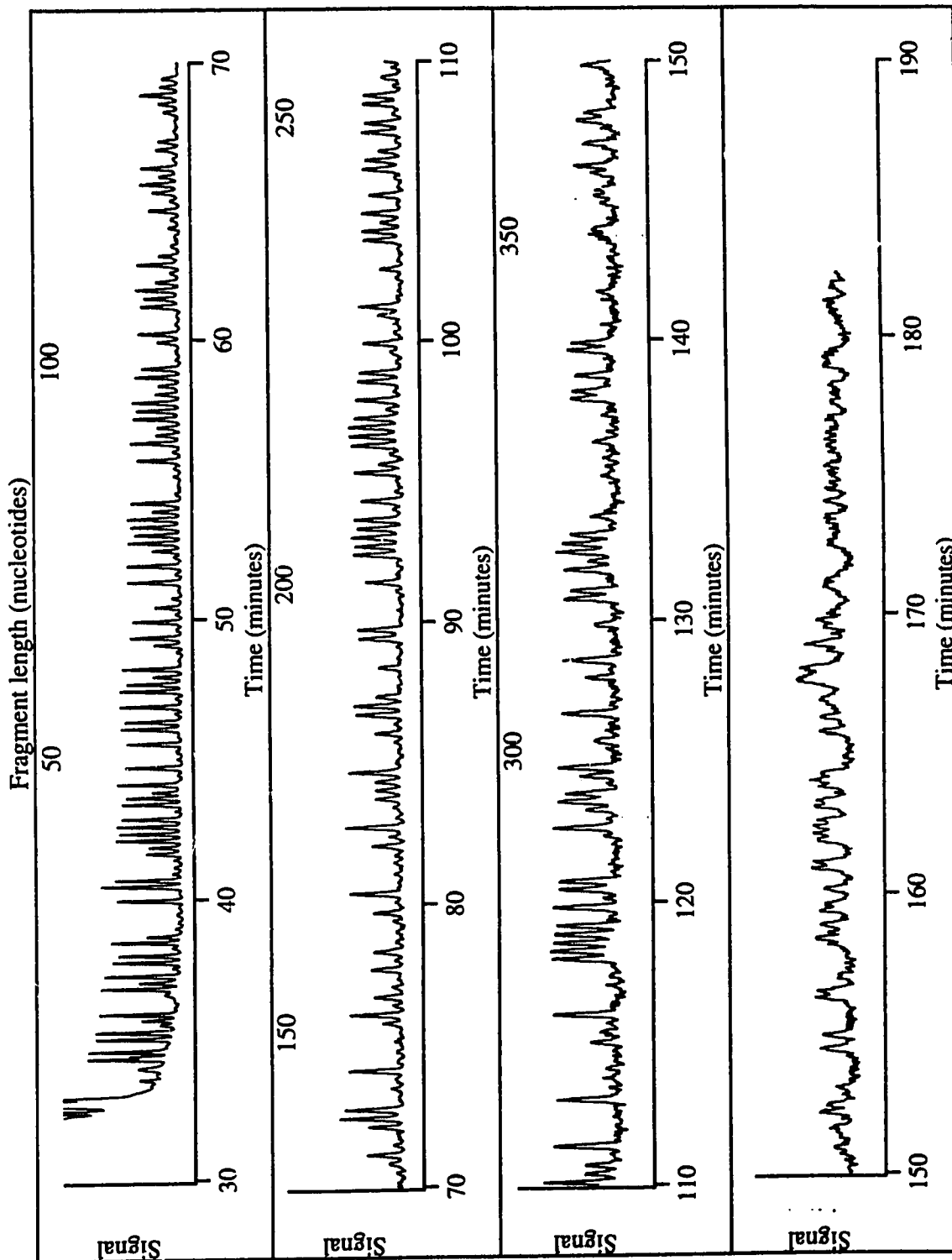


Figure 3.2b Expanded electropherogram of the A/C terminations of the MMTV sample

b.



c.

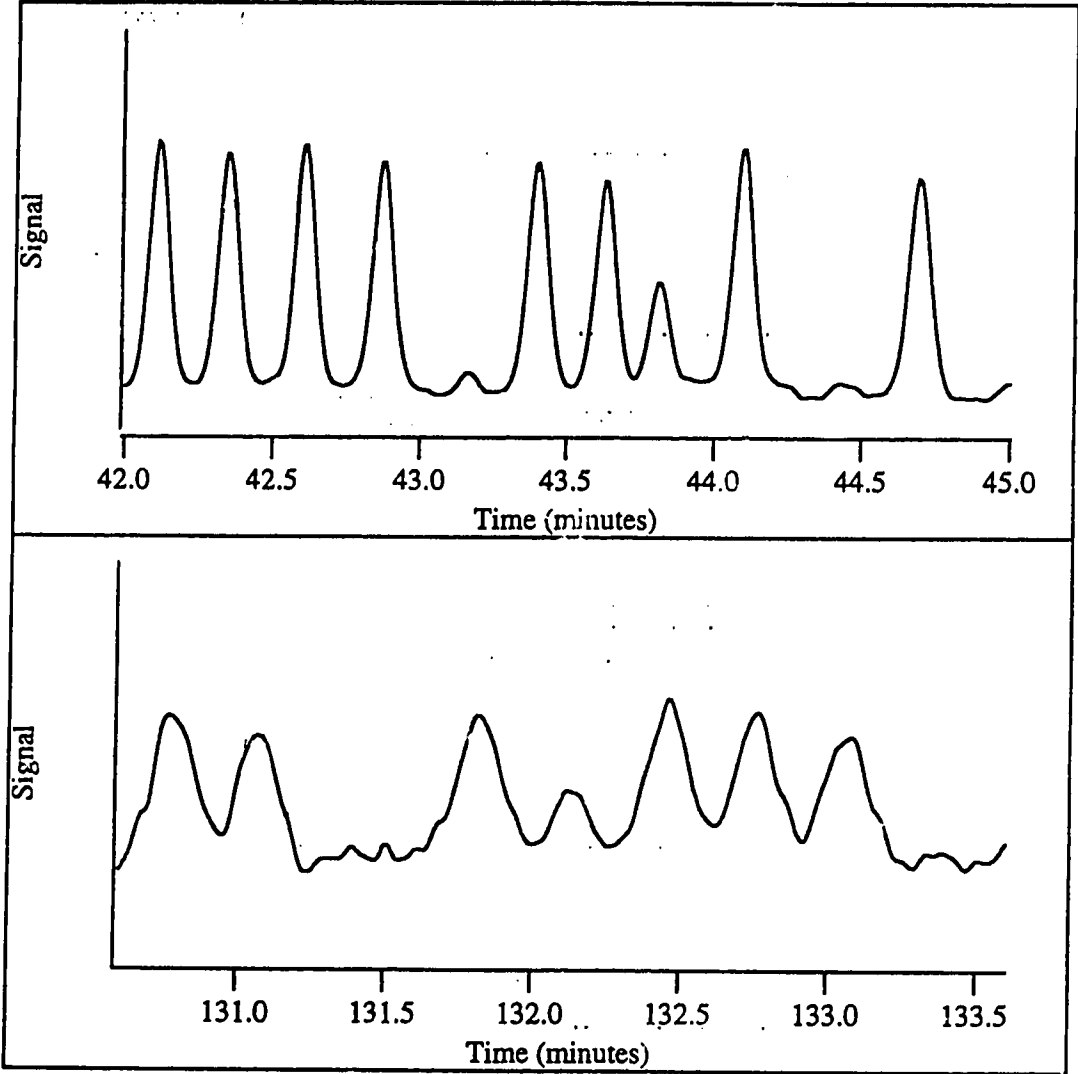


Figure 3.2c Fragments 40 to 50 (top panel) and 320 to 330 (bottom panel) nucleotides in length.

Figure 3.3a Electropherogram of a T and G terminated MMTV sample. Large peaks are T terminated fragments, small peaks are G terminated fragments.

a.

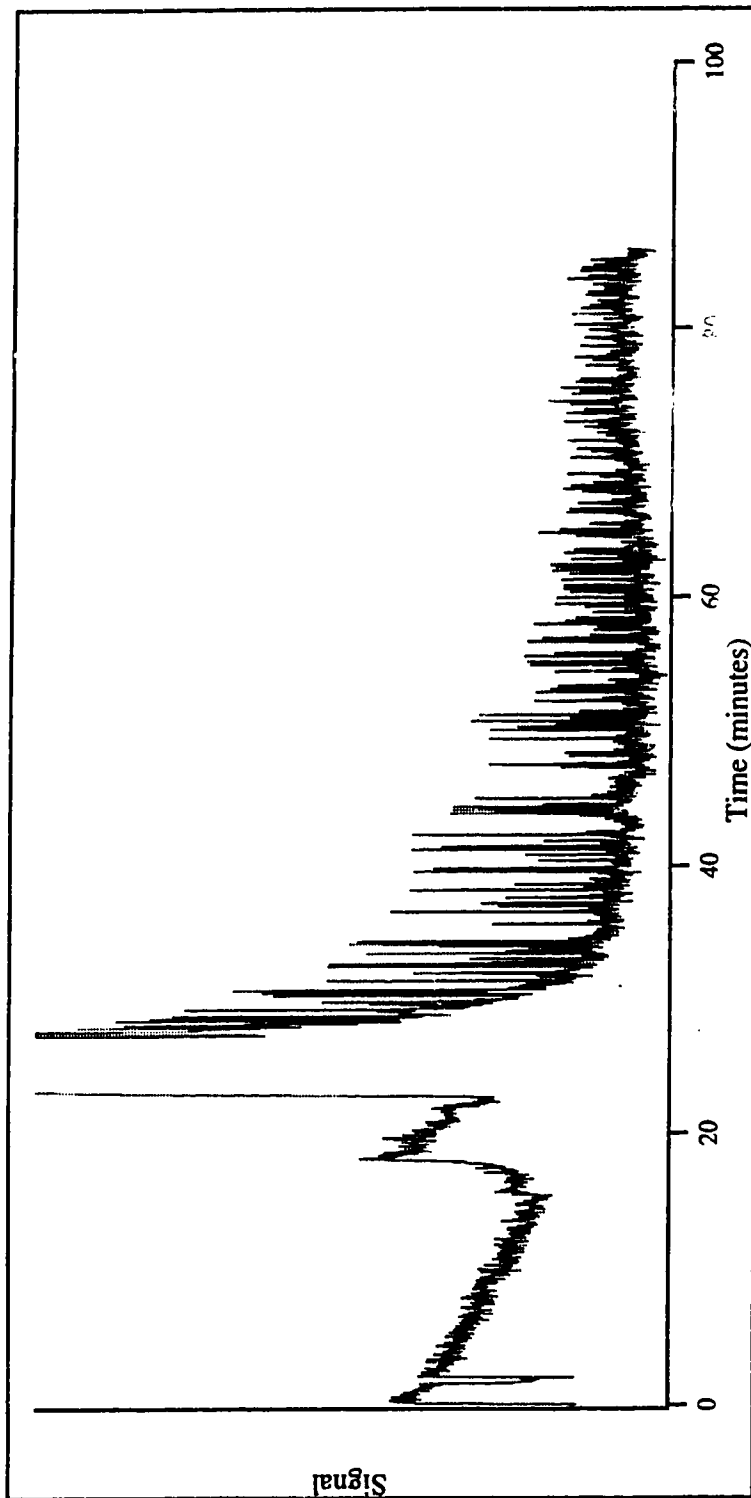
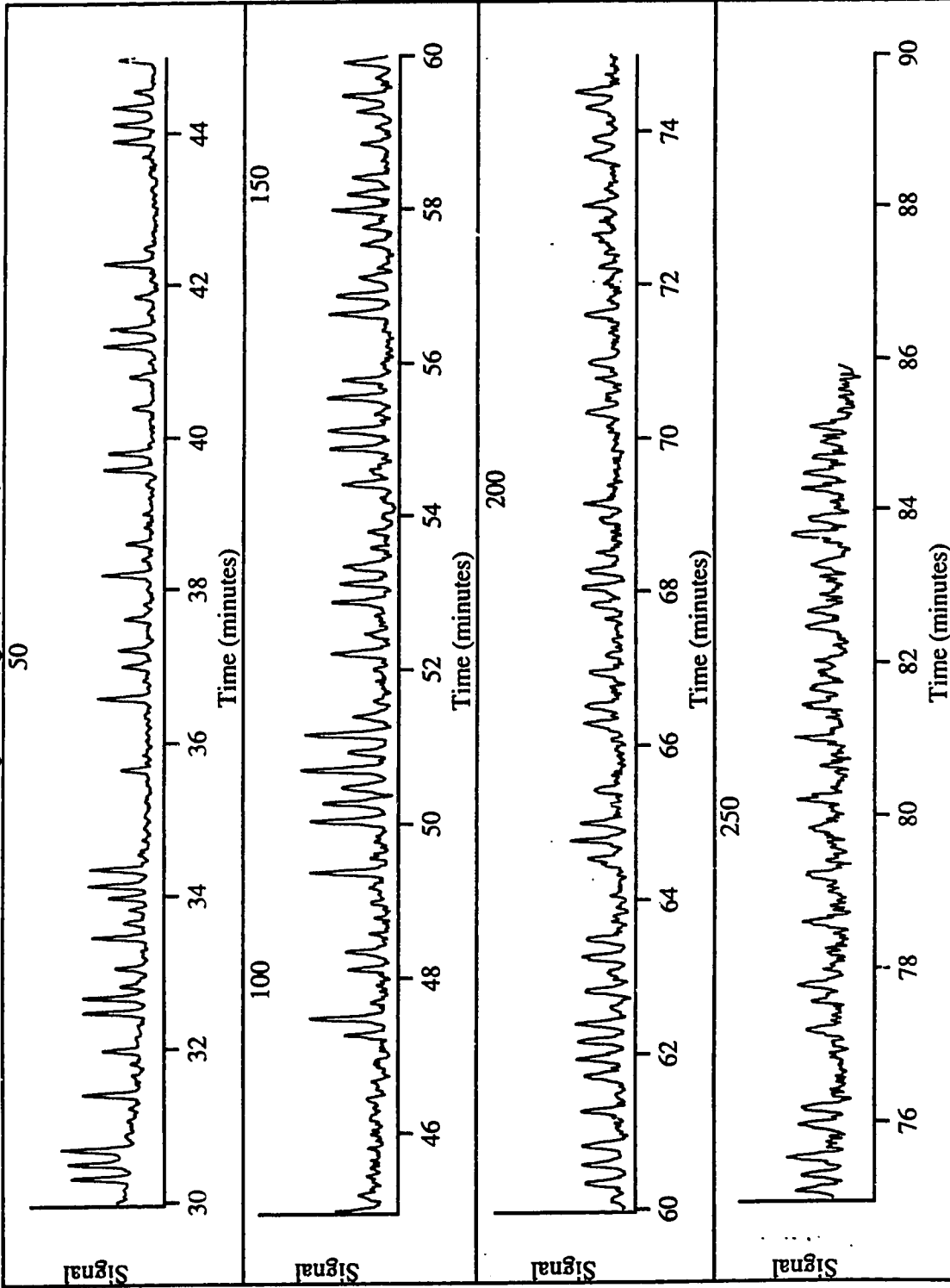


Figure 3.3b Expansion of the electropherogram of the T/G terminated MMTV sample

Fragment length (nucleotides)



b.

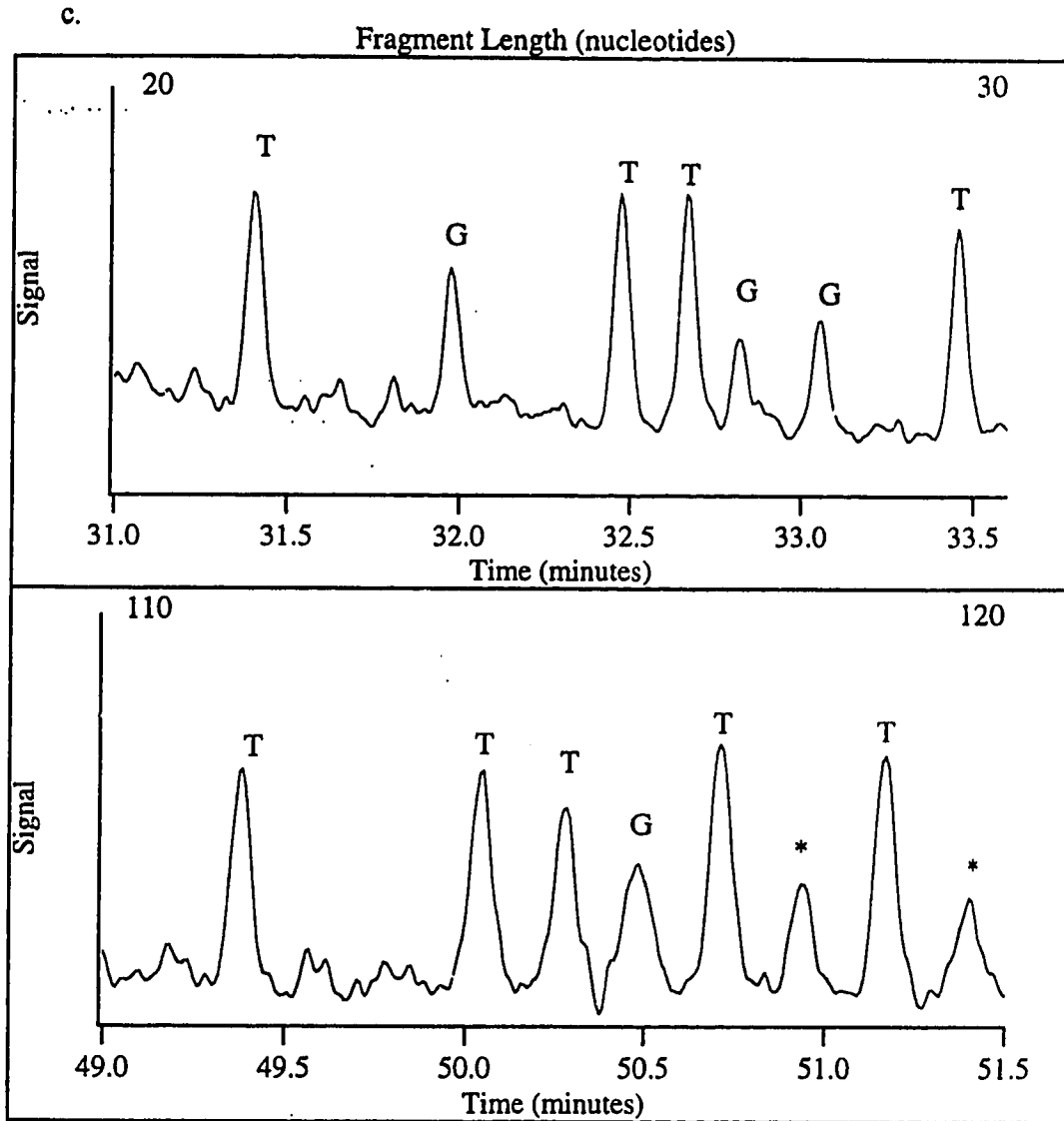
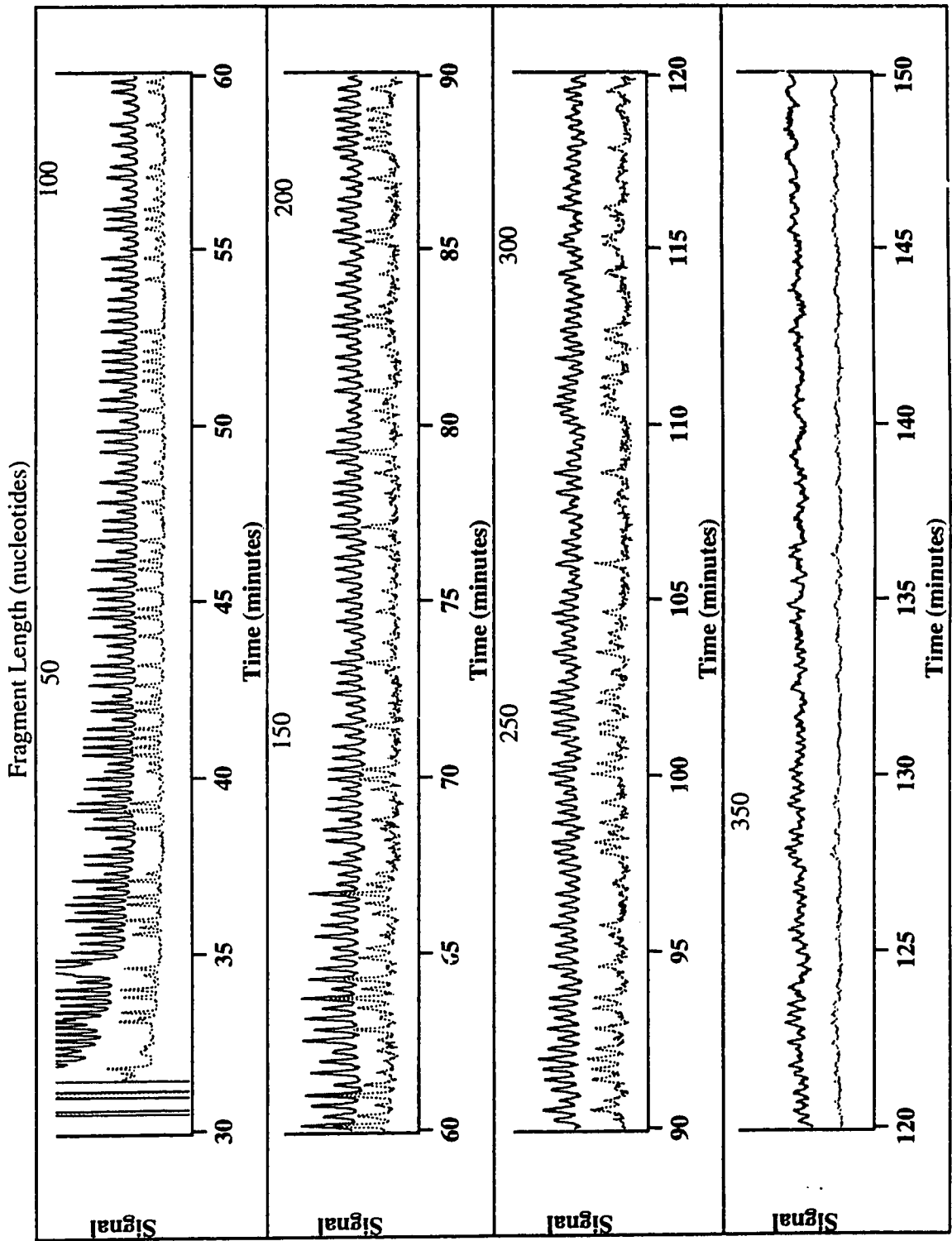


Figure 3.3c Fragments 20 to 30 (top panel) and 110 to 120 (bottom panel) nucleotides in length.

Figure 3.4 Electropherogram of MMTV sample with a peak height ratio of 2:1.
Peaks in both traces are T and G; large peaks are T, small peaks are G; Peaks in one trace only are A and C; large peaks are A, small peaks are C.



Samples sequenced with the 2:1 peak height ratio both with the FAM/JOE primer pair and with other pairs of primers (13) generated an average error rate of 3% for fragments of less than 340 bases in length. The errors are mostly due to poor height discrimination. To improve discrimination, a 3:1 peak height ratio was used. An electropherogram of an MMTV sequencing sample is shown in figure 3.5. The peak height ratios are 3:1. Peaks in both the solid and dashed traces are T and G; large peaks are T, small peaks are G. Peaks in the solid trace only are A and C; large peaks are A, small peaks are C. The sequence was determined to 400 nucleotides. The agreement of the determined sequence with the known sequence for MMTV was 99.5%. The errors were associated with fragments terminated by C. These fragments give small peaks in one channel only; the small peaks become difficult to interpret as the signal to noise ratio degrades late in the run.

To demonstrate the utility of the two colour peak height encoding method two additional samples were prepared. The first was prepared with M13mp18 control template DNA. The electropherogram is shown in figure 3.6. The sequence for M13mp18 was determined to 250 nucleotides with 97.6% agreement with the known sequence. The main sources of error were a compression at bases 65 to 68, and two G's that were missed at positions 211 and 213 due to poor signal to noise ratio and degrading resolution. Compressions occur in G-C rich portions of DNA when the DNA strand folds back and base pairs with itself. The resulting fragment migrates faster than expected, causing several peaks to elute very close together. The second sample was an insert of mouse cytokine DNA. The electropherogram is shown in figure 3.7. The sequence was determined to 300 nucleotides with 95% agreement with the known sequence. The errors were mainly due to low intensity peaks that were poorly resolved from high intensity peaks, and the compression at about 63 minutes.

Figure 3.5 Electropherogram of MMTV sample with a peak height ratio of 3:1.
Peaks in both traces are T and G; large peaks are T, small peaks are G; Peaks in one trace only are A and C; large peaks are A, small peaks are C.

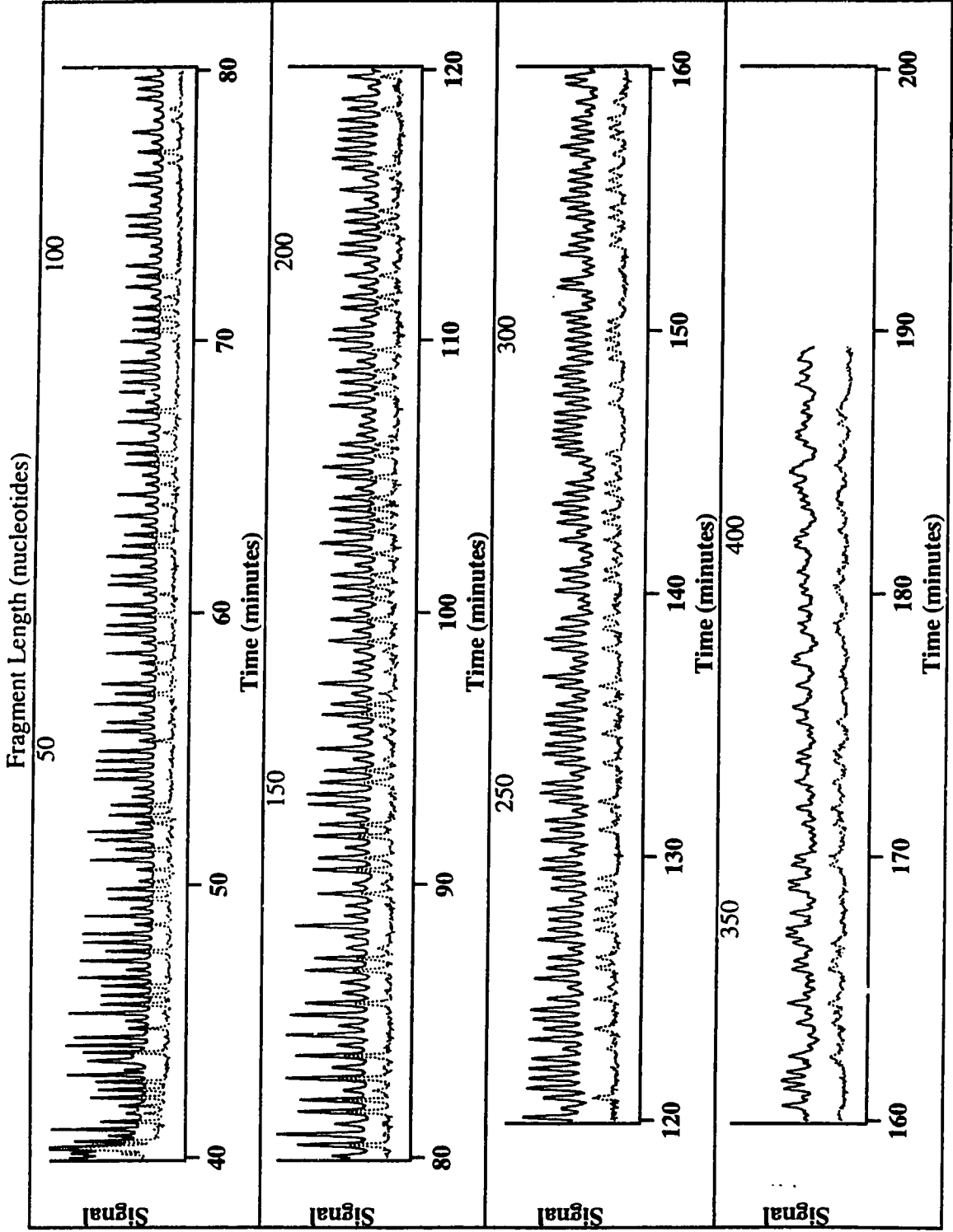


Figure 3.6 Electropherogram of M13mp18 sample. Peaks in both traces are T and G; large peaks are T, small peaks are G; Peaks in one trace only are A and C; large peaks are A, small peaks are C.

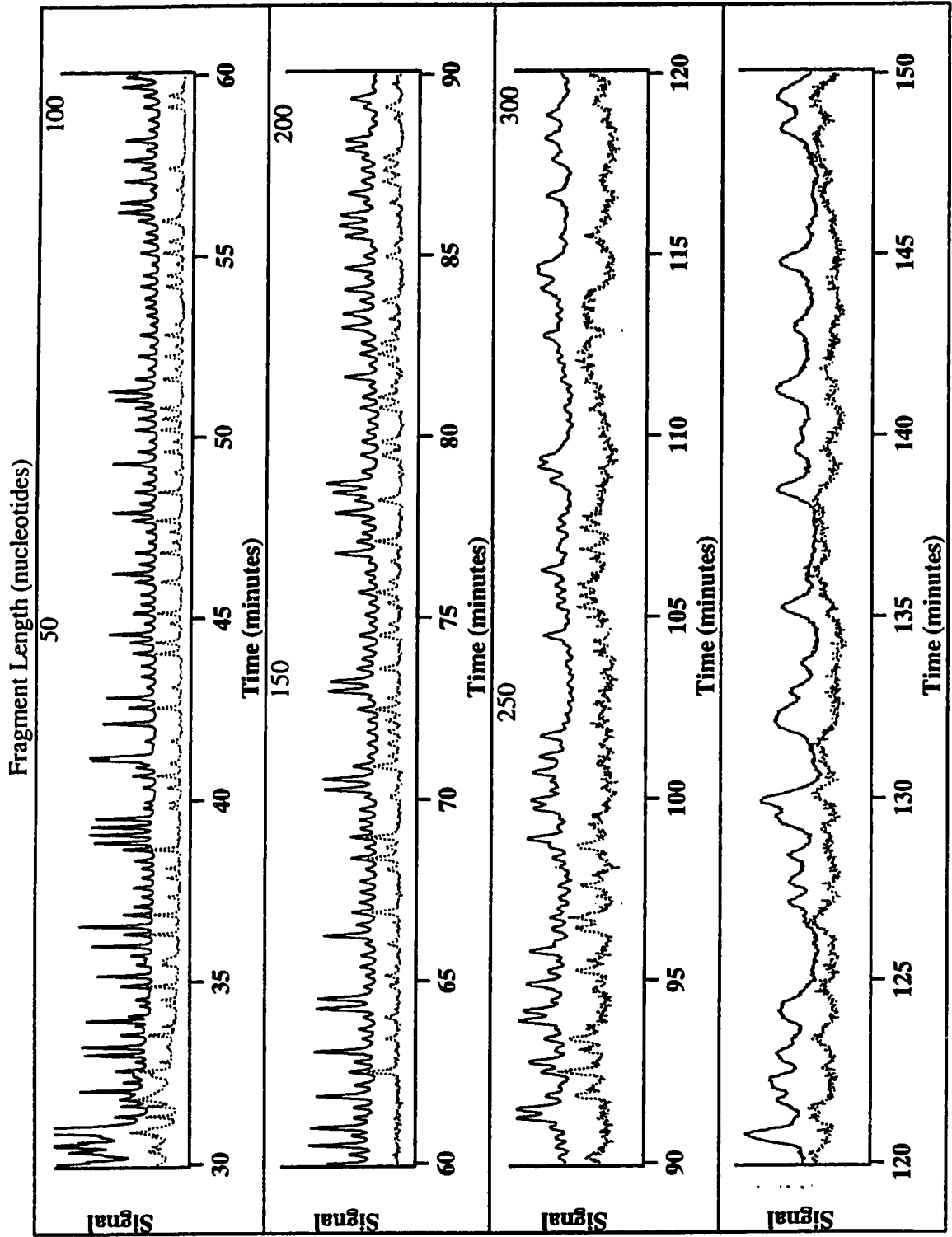
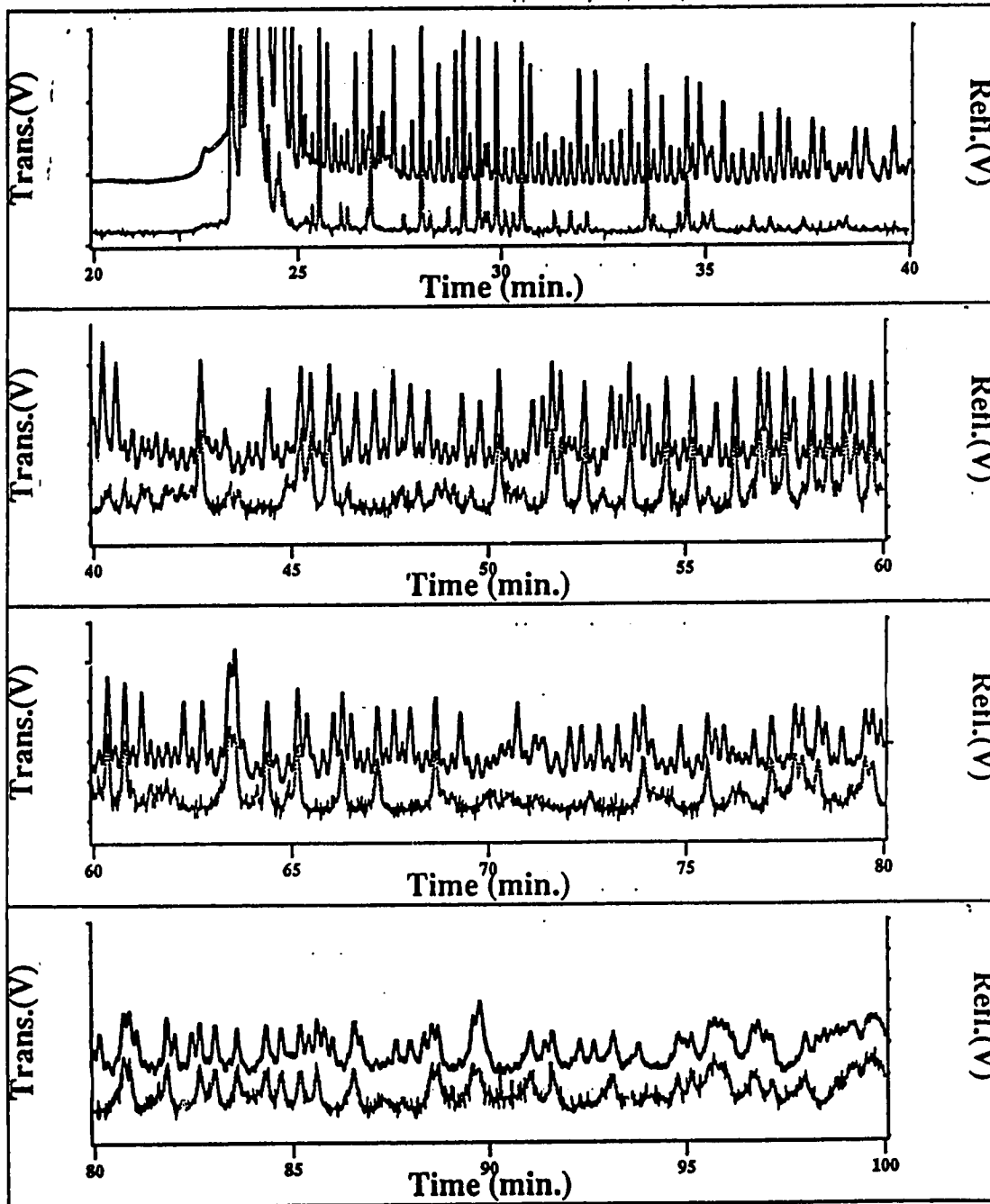


Figure 3.7 Electropherogram of mouse cytokine DNA. Peaks in both traces are T and G; large peaks are T, small peaks are G; Peaks in one trace only are A and C; large peaks are A, small peaks are C.



3.4 Conclusions

There is an optimum ratio of peak heights for this sequencing technique. If the ratio is near 1, then errors arise because of insufficient discrimination between peak heights. If the ratio is larger than 3, then errors are made late in the run for low amplitude peaks. It appears that a ratio of 2.5 to 3 is ideal in obtaining high sequencing accuracy for at least 400 bases. Improved signal-to-noise and computer algorithms will inevitably produce superior results for longer fragments.

Sequencing accuracy was determined by visual inspection of the data followed by comparison with the known sequence. The best data was obtained with an A-T rich sample that tends not to form compressions; sequencing accuracy exceeded 99% for fragments up to 400 bases in length.

The two-colour peak height encoded system produces accuracy equal to or greater than conventional fluorescence based sequencing (14). However, only two fluorescently labelled primers are required for the sequencing reaction, compared with four labels for the ABI sequencer. As a result, the two-colour peak height encoded system could be cost effective for primer walking experiments, particularly when the rapid throughput of capillary electrophoresis outweighs the disadvantages of preparing two labelled primers. The two-colour sequencing protocol suffers from one significant disadvantage: the technique requires use of T7 polymerase. Until a thermally stable polymerase is discovered that produces uniform incorporation of dideoxynucleosides, cycle sequencing and PCR-based sequencing will remain impractical with the peak-height encoded sequencing techniques.

Additional systems have been designed and described for use with the following pairs of Applied Biosystems primer: FAM-TAMRA, JOE-ROX, JOE-TAMRA, TAMRA-ROX and FAM-ROX(13). Any of these reaction pairs, including

the FAM-JOE pair described in this chapter, could be separated on the ABI sequencer. The FAM-JOE pair could also be separated with the DuPont Genesis sequencer. The peak-height encoded technique may be employed on commercial sequencers with appropriate software modification.

References

1. S. Tabor, C.C. Richardson, *Proc. Natl. Acad. Sci. USA*, **86** (1989), 4076.
2. S. Tabor, C.C. Richardson, *J. Biol. Chem.*, **265** (1990), 8322.
3. W. Ansorge, J. Zimmermann, C. Schwager, J. Stegemann, H. Erfle, H. Voss, *Nucleic Acids Res.*, **18** (1990), 3419.
4. H. Swerdlow, R. Gesteland, *Nucleic Acids Res.*, **18** (1990), 141.
5. H. Drossman, J.A. Luckey, A.J. Kostichka, J. D'Cunha, L.M. Smith, *Anal. Chem.*, **62** (1990), 900.
6. A.S. Cohen, D.R. Najarian, B.L. Karger, *J. Chromatogr.*, **516** (1990), 49.
7. H.P. Swerdlow, S. Wu, H.R. Harke, N.J. Dovichi, *J. Chromatogr.*, **516** (1990), 61.
8. J.A. Luckey, H. Drossman, A.J. Kostichka, D.A. Mead, J. D'Cunha, T.B. Norris, L.M. Smith, *Nucleic Acids Res.*, **18** (1990), 4417.
9. D.Y. Chen, H.P. Swerdlow, H.R. Harke, J.Z. Zhang, N.J. Dovichi, *J. Chromatogr.*, **559** (1991), 237.
10. A.E. Karger, J.M. Harris, R.F. Gesteland, *Nucleic Acids Res.*, **19** (1991), 4955.
11. H.P. Swerdlow, J.Z. Zhang, D.Y. Chen, H.R. Harke, R. Grey, S. Wu, C. Fuller, N.J. Dovichi, *Anal. Chem.*, **63** (1991), 2835.
12. H.R. Harke, S. Bay, J.Z. Zhang, M.J. Rocheleau, N.J. Dovichi, *J. Chromatogr.*, **608** (1992), 143.
13. D.Y. Chen, H.R. Harke, N.J. Dovichi, *Nucl. Acids Res.*, **20** (1992), 4873.
14. T. Hunkapillar, R.J. Kaiser, B.F. Koop, L. Hood, *Science*, **254** (1991), 59.

CHAPTER 4

Accuracy of Two-Colour Peak Height Encoded DNA Sequencing by Capillary Gel Electrophoresis and Laser-Induced Fluorescence³.

4.1 Introduction

Capillary gel electrophoresis (CGE) is a powerful high speed approach for the separation of biological macromolecules. Sequence analysis of deoxyribonucleic acid (DNA) is one area where the application of CGE will be useful. The Human Genome Project has set the goal of sequencing the human genome, approximately 3 billion base pairs containing 50 000 to 100 000 genes, in the next 15 years .

Conventional sequencing techniques are slow and labour intensive (1-4). By comparison, a six to twelve hour separation by slab gel electrophoresis can be completed in minutes by CGE (5-7), representing an order of magnitude improvement in separation speed. In addition to improvements in sequencing rate, CGE also produces higher resolution and separation efficiency (5-7).

Automated DNA sequencers with fluorescent detection are generally based on one of two schemes. The Pharmacia ALF instrument uses one fluorescent primer for all four reactions the products of which are then separated on four adjacent lanes of a gel (3). The Applied Biosystems instrument uses four different fluorescent labels where one label is associated with each dideoxynucleotide chain terminator (ddNTP) (4,8). Ambiguities arise in both systems associated with compressions and false peaks arising from false priming and exonuclease activity. Ambiguities in the ALF system arise when mobilities of fragments in adjacent lanes differ due to thermal gradients.

³A version of this chapter has been published in *Advances in DNA Sequencing*, SPIE, Vol. 1891 (1993), p8.

Ambiguities in the ABI data arise due to poor signal-to-noise leading to inaccurate identification of the fluorescent label. Both systems produce reasonably high accuracy (>97%) (9-10) for the first 300 to 450 bases of sequence.

Capillary electrophoresis systems are not limited to commercially available sequencing strategies (11). Chapter 3 details a strategy which requires two dyes for DNA sequencing. This technique is a modification of the peak height encoded method reported by Richardson and Tabor (12), and relies on the uniform incorporation of dideoxynucleosides by T7 polymerase in the presence of manganese ions (13). Two commercially available fluorescently labelled primers are used to generate two separate samples. The first labelled primer is used in a reaction with template DNA, deoxynucleoside triphosphates (dNTPs), dideoxyadenosine triphosphate (ddATP) and dideoxycytosine triphosphate (ddCTP). The ddATP and ddCTP are added in a 3:1 ratio (A:C). The second primer is used in a reaction with dideoxythymidine triphosphate (ddTTP) and dideoxyguanosine triphosphate (ddGTP) added in a 3:1 ratio (T:G). The reaction products are combined and separated in a single gel-filled capillary. The sequence is determined based on the combination of amplitude and spectral information. Big peaks associated with the first dye are due to A, small peaks are due to C. Big peaks associated with the second dye are due to T, small peaks are due to G.

This chapter provides a direct comparison of DNA sequence obtained by CGE with that obtained with the Pharmacia ALF sequencer. The sample templates are DNA segments from the malaria genome. Sequences obtained by each method are compared to each other and to the final consensus sequence in order to obtain an evaluation of the accuracy of the data generated by CGE.

4.2 Experimental

4.2.1 Instrument Design

The instrument is described in detail in Section 2.2 and the detector is shown in figure 3.1. A Plexiglas box, equipped with a safety interlock, holds the injection end of the capillary. DNA fragments are forced through the capillary by the application of a negative high voltage at the injection end. The other end of the capillary is inserted into the flow chamber of the sheath flow cuvette held at ground potential. The cuvette has a 200 μm square flow chamber.

Fluorescence is excited with an argon ion laser operating at 488 nm with a power of 30 mW, and is collected at right angles with a microscope objective and imaged onto a pinhole. Transmitted light illuminates a dichroic beam splitter that transmits light at wavelengths greater than 550 nm and reflects light at wavelengths less than 550 nm. The fluorescence in each channel is passed through an appropriate band pass filter to a photomultiplier tube (PMT). Current from each PMT is conditioned with a 0.5 sec low pass filter, digitised and recorded with a Macintosh IIsi computer.

Polyimide coated fused silica capillaries are used with 50 μm inner diameter, 190 μm outer diameter and typically 40 cm in length.

4.2.2 Gel Preparation

The gels were prepared from 5 ml aliquots of a 4%T 5%C mixture of acrylamide and bisacrylamide (N,N'-methylenebisacrylamide) in 1 X TBE and 7 M urea. The solution was degassed and polymerisation was initiated by the addition of 2 μl TEMED and 20

μl of 10% (w/v) ammonium persulphate. Before injection, the capillary was silanised for approximately 2 cm from the detection end with γ -methacrylpropyltrimethoxysilane. Silanisation covalently binds the gel to the capillary and prevents movement of gel into the detection cuvette. The acrylamide solution was mixed and cooled for a few minutes on ice to prevent excessively rapid polymerisation, and then injected into the capillary with a syringe. The polymerisation of the gel was complete in 30 minutes but gels were typically stored overnight before use.

4.2.3 Sample Preparation

Six samples of single stranded malaria DNA were used. The concentration of template DNA was approximately 0.125 mg/ml for clones 1 to 5. The sequencing procedure was based on the protocol included in the Sequenase Dye-Primer Sequencing Kit (manufactured by United States Biochemical for Applied Biosystems; no longer commercially available). For clone 1, 1.25 μg of template DNA was added to 1.6 pmol JOE labelled T3 primer (0.4 μM), 2.5 μl 10X MOPS buffer (400 mM MOPS, pH 7.5, 500 mM NaCl, 100 mM MgCl_2), 2.5 μl Mn solution (50 mM MnCl_2 , 150 mM sodium isocitrate), and water to a volume of 18 μl . The primer and template were annealed by heating the mixture to 65°C for 2 minutes, followed by slow cooling to room temperature. Four microlitres of the A/C termination mix (1 mM in each of dATP, dCTP, dGTP, dTTP; 2.5 μM ddATP, 0.8 μM ddCTP) were added. The ratio of the dideoxynucleosides in the A/C termination mix was adjusted to yield a peak height ratio of 3:1 for A to C. The reaction mixture was then preheated at 37°C for two minutes, and then 6.5 units of Sequenase Version 2.0 and 0.006 units of pyrophosphatase were added. The mixture was incubated at 37°C for 10 minutes after which 12 μl of stop/salt solution (20 mM EDTA, 1 M sodium acetate, pH 8.0) was

added. The DNA was precipitated by addition of 120 μ l of 98% ethanol. The sample was placed at -20°C for at least 20 minutes, but typically overnight. The sample was spun for 20 min in a centrifuge and the supernatant was removed. The sample was then washed with 300 μ l of ice cold 80% ethanol, spun again for 20 minutes and dried. It was then resuspended in 4 μ L of 49:1 formamide:0.5M EDTA, pH 8.0. The same conditions were used with FAM labelled primer to yield a peak height ratio of 3:1 for T to G. Two microlitre aliquots were taken from each resuspended sample and mixed.

The reaction procedure was the same for the other samples except the amount of sample template and primer were varied. For clone 2, 0.8 pmol of primer was annealed to 0.6 μ g of template 2. For clone 3, 0.8 pmol of JOE primer was annealed to 0.6 μ g of template 3 and 0.4 pmol of FAM primer was annealed to 0.3 μ g of template 3. For clone 4, 0.8 pmol of JOE primer was annealed to 0.6 μ g of template 4 and 0.4 pmol FAM primer was annealed to 0.3 μ g of template 4. For clone 5, 0.8 pmol of primer was annealed to 0.4 μ g of template 6. A second sample, was prepared from clone 2 by annealing 0.8 pmol of JOE primer with 0.6 mg of template 2, and 0.4 pmol of FAM primer with 0.3 mg of template 2. The termination mixes were combined to give peak height ratios of 3:1 C:A for the JOE labelled fragments, and 3:1 G:T for the FAM labelled fragments.

4.2.4 Electrophoresis

Before injection, each sample was denatured at 95 C for 2 minutes. Clones 1 through 4 were injected for 30 seconds at 200V/cm. Clone 5 was injected for 30 seconds at 175 V/cm. For all other samples electrophoresis was carried out at 200 V/cm. The sheath flow stream was 1X TBE at a flow rate of 0.16 ml/hour. The runs were performed at room temperature, 20°C .

The DNA sequence was read from the smoothed electrophoresis data by eye. The sequence was then compared with the sequence generated on the Pharmacia ALF instrument as well as with the consensus sequence.

4.2.5 ALF Sequencing

The same templates as above were sequenced using the standard Pharmacia ALF protocol on 0.5 mm thick 6%T, 5%C polyacrylamide, 7 M urea slab gels with 0.6X TBE running buffer. Samples were prepared using procedure A of Pharmacia's AutoRead Sequencing Kit.

4.3 Results And Discussion

Sequencing electropherograms from clones 1 through 5 are shown in figures 4.1 to 4.5. The A-T rich malaria DNA produced runs that were free of compressions and had good signal-to-noise ratios. The runs generated by capillary gel electrophoresis were compared: 1. with individual slab gel runs of the same template DNA done on the Pharmacia ALF sequencer and 2. with the consensus sequence which was compiled from sequence of overlapping clones and the sequence of both strands of the malaria DNA.

4.3.1 Sequence Accuracy

The sequences were read and compared with the consensus sequences for the malaria genome. A consensus sequence is determined by sequencing overlapping DNA clones, and by sequencing both complementary strands of DNA. In this way a reliable DNA sequence is obtained.

Errors occur when there is a discrepancy between an individual run and the consensus sequence. Ambiguities occur in sequences where the identity of a particular nucleotide is unclear. When the electropherograms are inspected by eye, a reasonable guess may be attempted. The accuracy of a sequence is the percentage of nucleotides that agree with the consensus sequence. For clone 1 a sequence 463 bases long was read from the capillary electropherogram. The sequence contained 5 errors and the accuracy is 98.9%. The ALF sequencer generated a sequence 435 bases long. There were 3 errors or ambiguities in this sequence and the accuracy is 99.3%. The accuracy results for the CGE sequences are summarized in Table 4.1. The accuracies range from 97.3% to 99.8%. The lowest accuracy was obtained for clone 3. Clone 3 was the poorest quality template of the five, resulting in the lowest accuracy and the shortest run (300 bases). The other four clones all gave accuracy values of 98% or higher. It is interesting to note that one error in the consensus sequence was discovered when the original commercial data were reinspected and the capillary gel data were confirmed.

The errors consisted mainly of problems distinguishing between G and C late in the CGE runs. In this labelling scheme, G and C are the lowest intensity peaks in the 3:1 ratio. Lower signal-to-noise ratio later in the run makes it difficult to distinguish between the two bases. Also, as the peak resolution decreases, the low intensity peaks are obscured and missed, as adjacent high intensity peaks tend to overshadow the

Figure 4.1 a. Electropherogram of malaria DNA clone 1. FAM label gives signal in both the solid and dashed traces, T>G; JOE label gives signal in the solid trace only, A>C. Electrophoresis was carried out at 200 V/cm.

a.

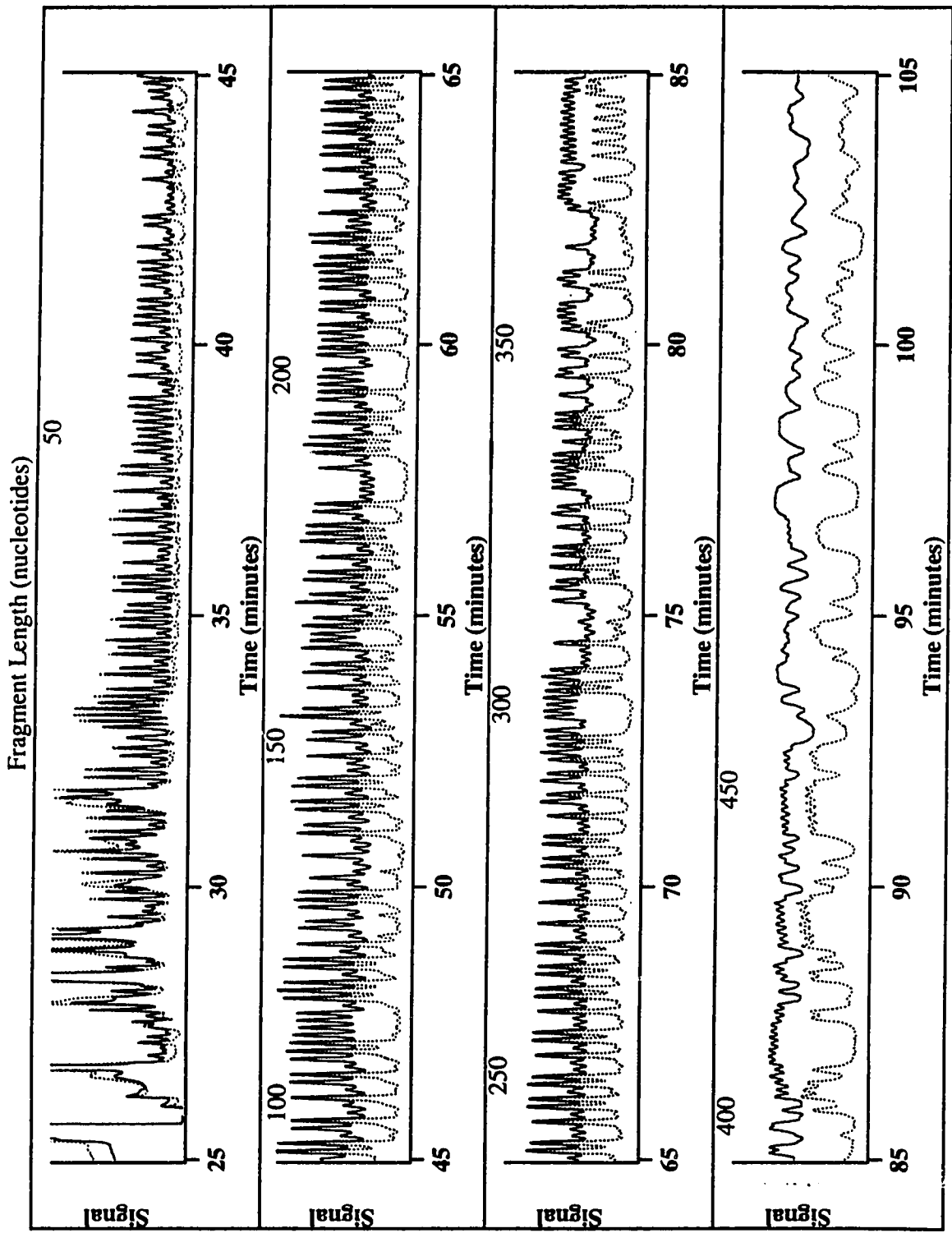


Figure 4.1 b. Expansion of two regions of the electropherogram showing fragments ranging from 100 to 150 and 350 to 400 bases in length. The corresponding sequence is printed at the top of each frame.

b.

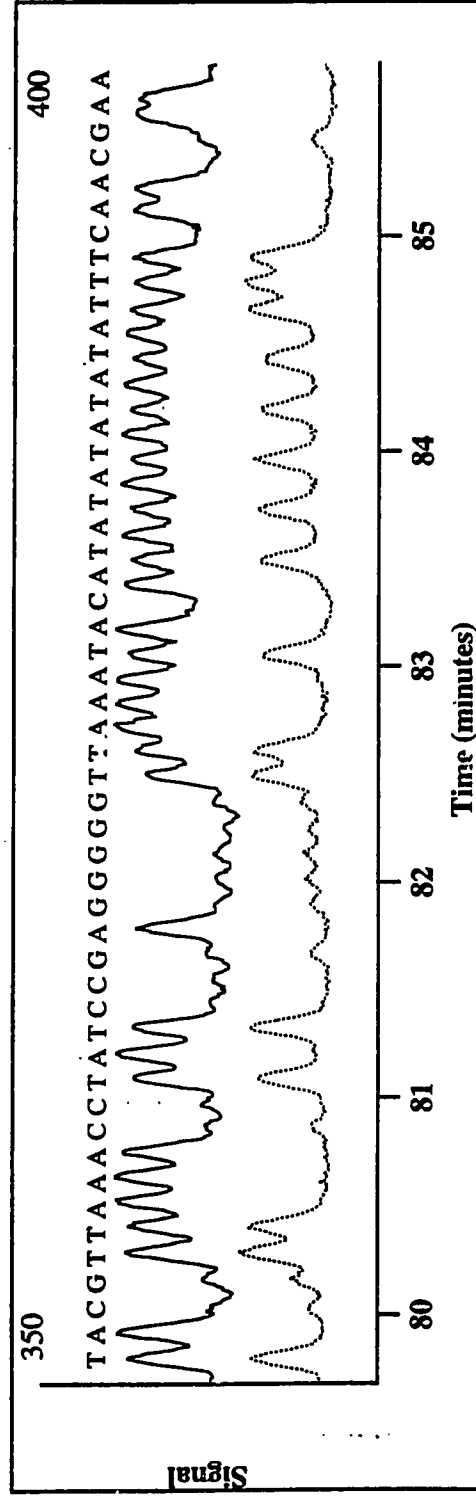
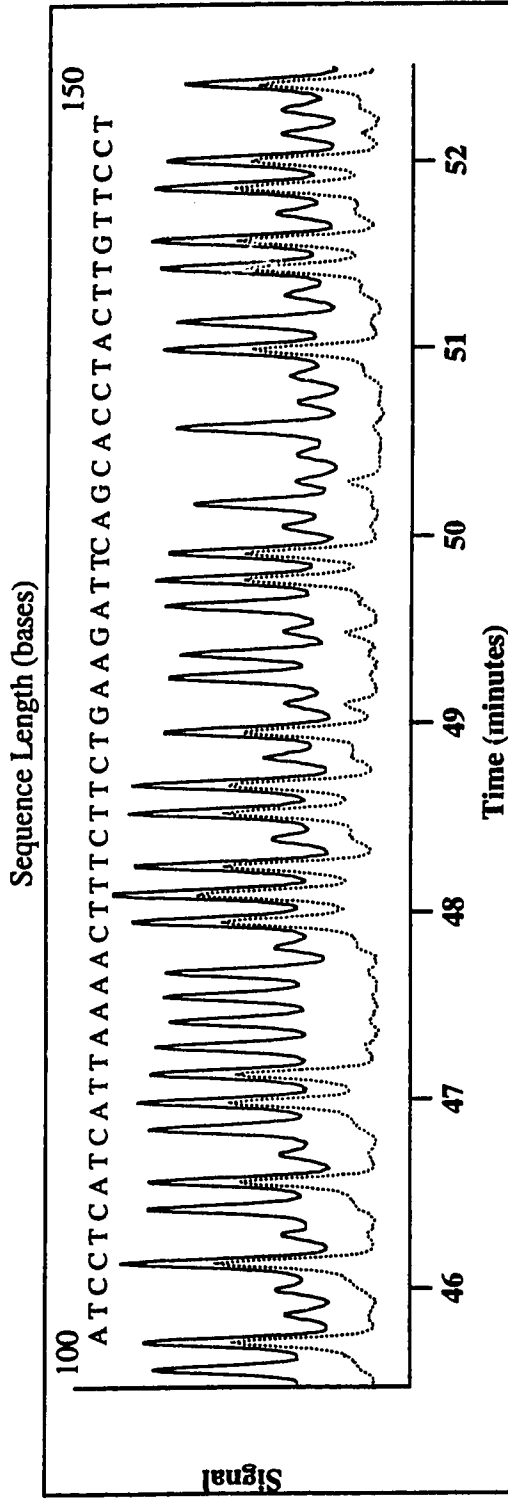


Figure 4.2 a. Electropherogram of malaria DNA clone 2. FAM label gives signal in both the solid and dashed traced, T>G; JOE label gives signal in the solid trace only, A>C. Electrophoresis was carried out at 200 V/cm.

a.

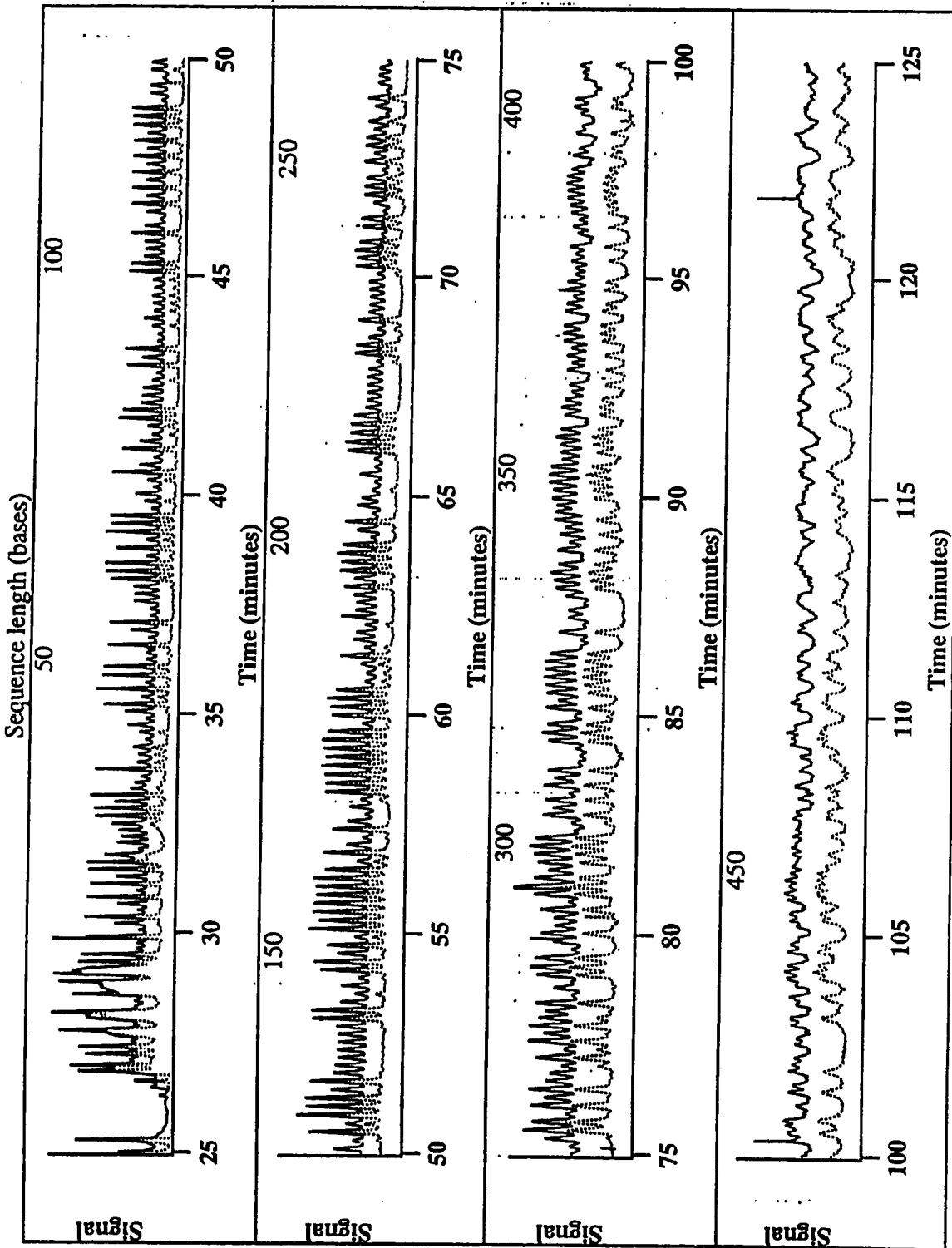


Figure 4.2 **b. Electropherogram of malaria DNA clone 2. FAM label gives signal in both traces, G>T; JOE in the solid trace only, C>A. Electrophoresis was carried out at 200 V/cm.**

b.

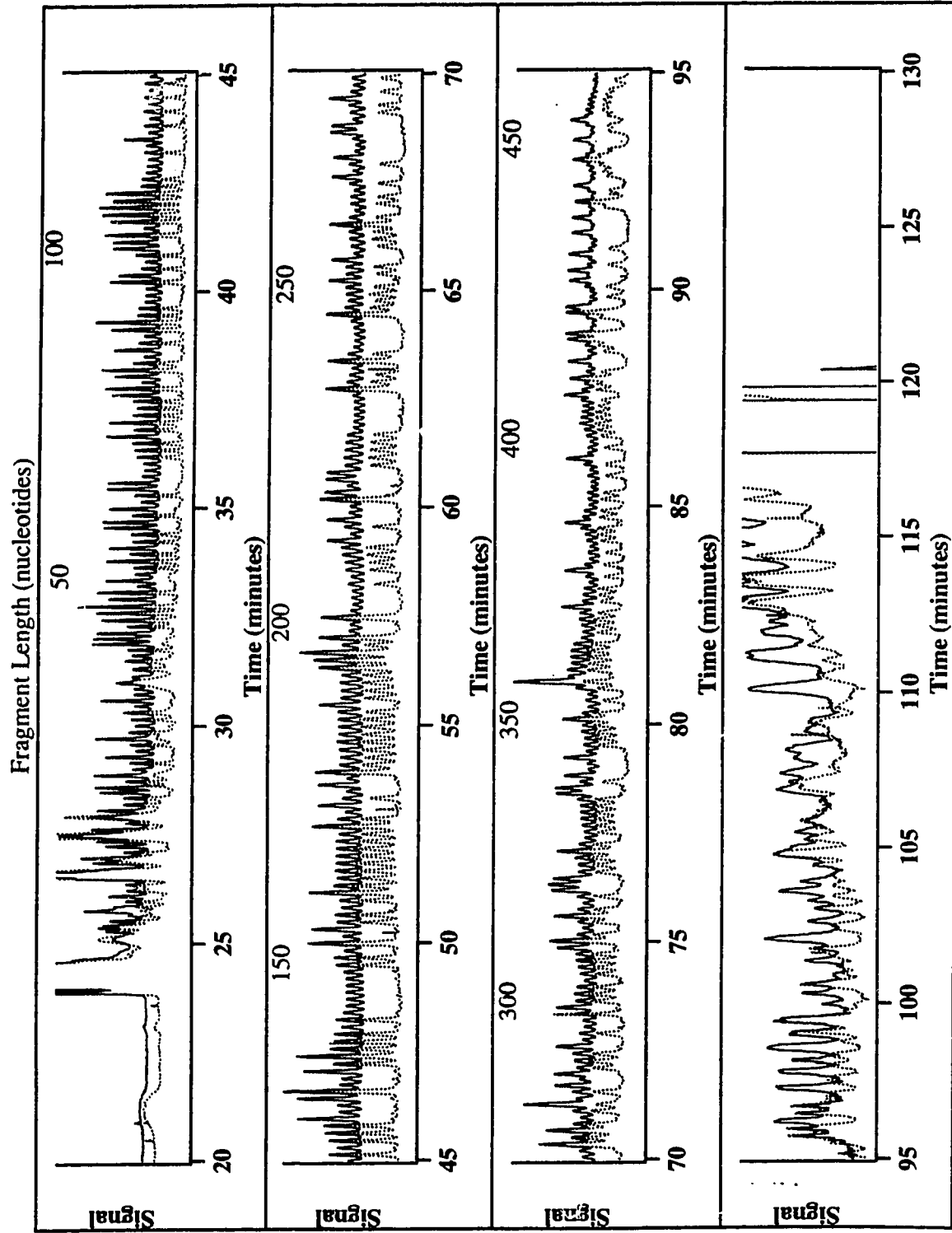


Figure 4.3 Electropherogram of malaria DNA clone 3. FAM label gives signal in both the solid and dashed traces, T>G; JOE label gives signal in the solid trace only, A>C. Electrophoresis was carried out at 200 V/cm.

Fragment Length (nucleotides)

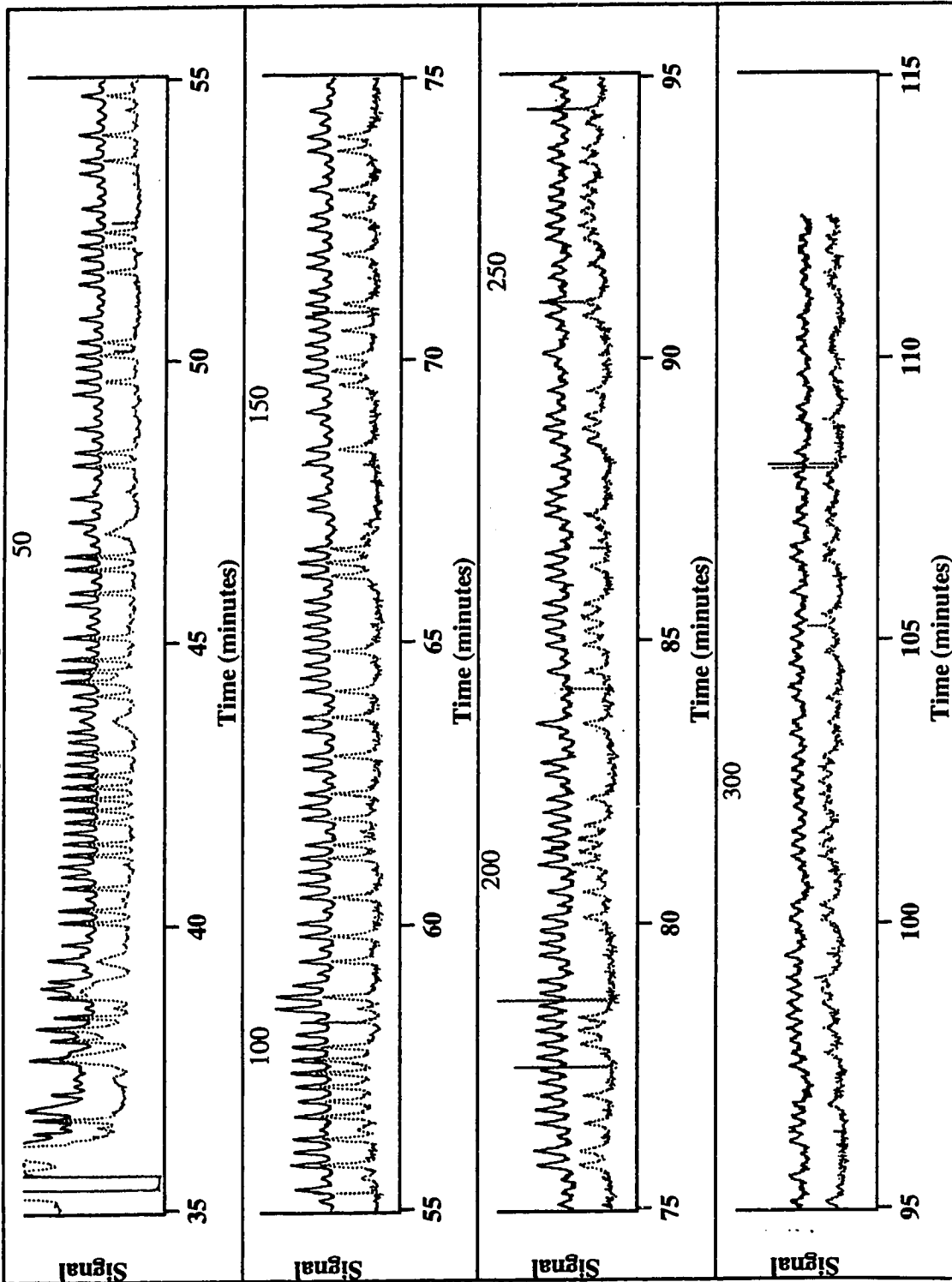


Figure 4.4 Electropherogram of malaria DNA clone 4. FAM label gives signal in both the solid and dashed traces, T>G; JOE label gives signal in the solid trace only, A>C. Electrophoresis was carried out at 200 V/cm.

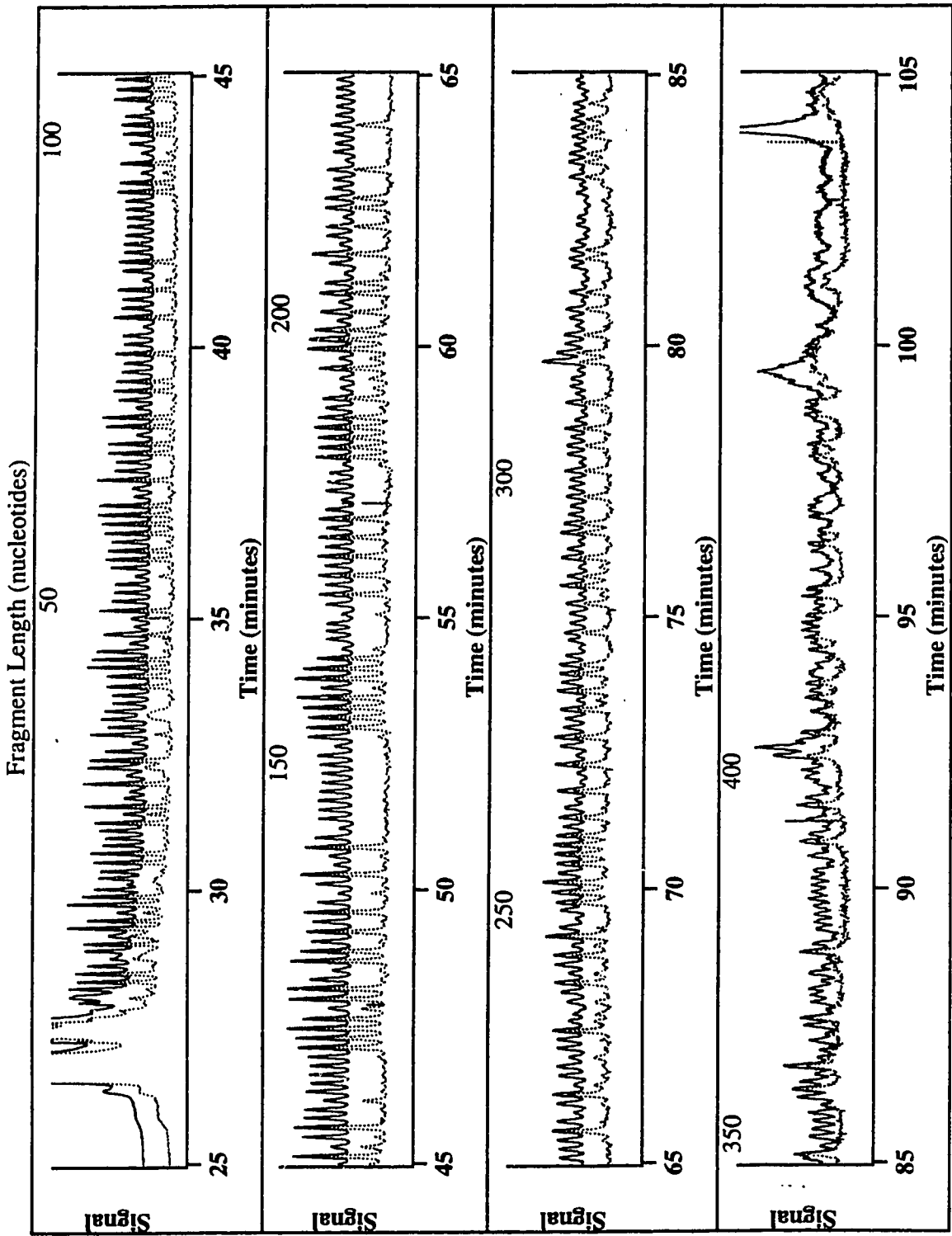
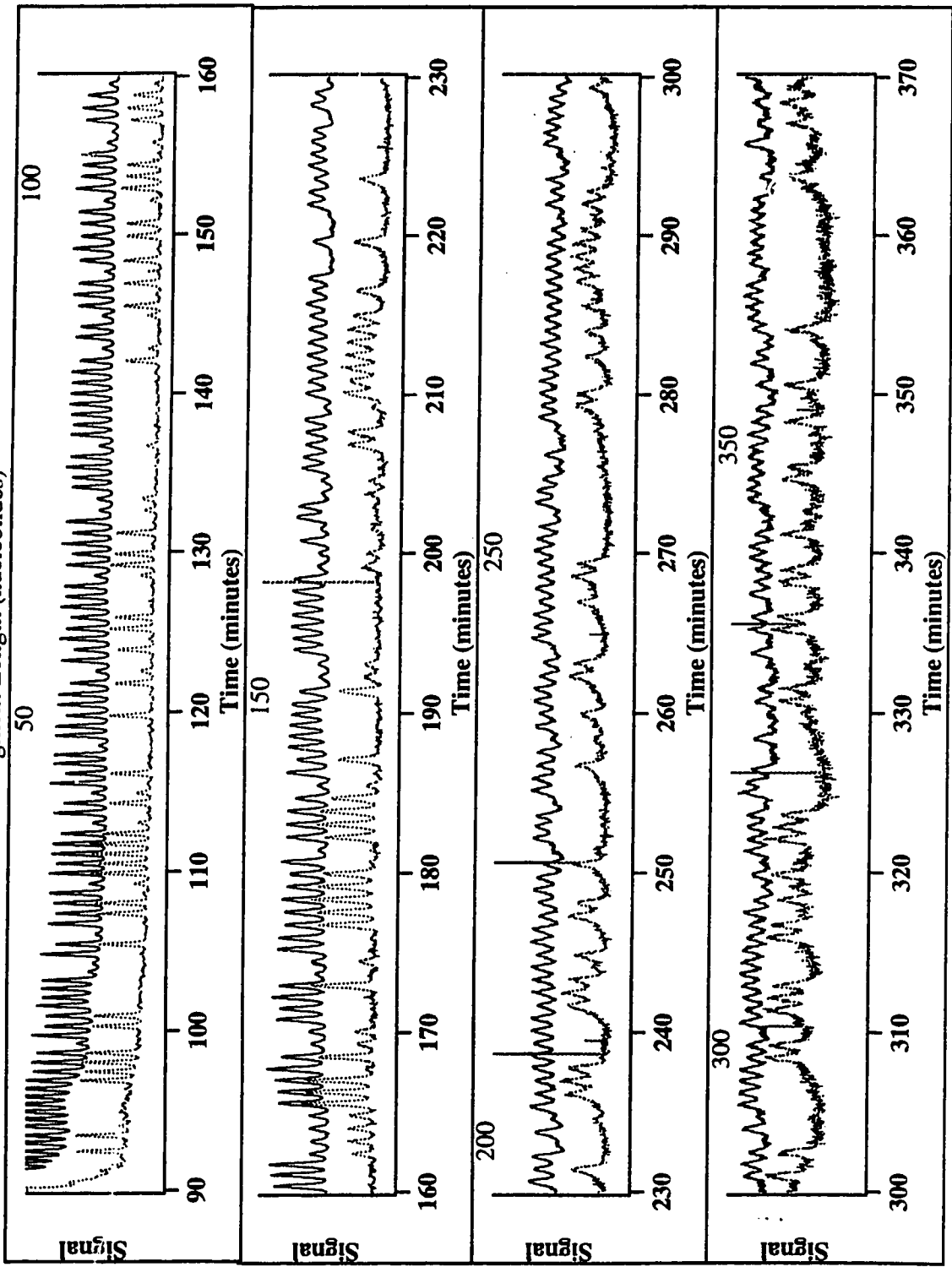


Figure 4.5 Electropherogram of malaria DNA clone 5. FAM label gives signal in both the solid and dashed traces, T>G; JOE label gives signal in the solid trace only, A>C. Electrophoresis was carried out at 175 V/cm.

Fragment Length (nucleotides)



small peaks. Sixty-three percent of all the errors occurred for fragments longer than 300 bases. If sample 3 is excluded (since the sequence only extends to about 300 bases) then 76% of the errors occurred for fragments longer than 300 bases and 52% for fragments longer than 400 bases. Clearly the sequence is most reliable up to 400 bases. This same trend has been reported by Koop et al. (14) for sequences generated with the ABI automated sequencer.

Table 4.1 Length and Accuracy of DNA sequence determined by CGE and automated slab gel electrophoresis

	Capillary Gel Electrophoresis		Pharmacia ALF Sequencer	
	Length of Run (bases)	Accuracy (%)	Length of Run (bases)	Accuracy (%)
Clone 1	463	98.9	435	99.3
Clone 2	450	98.4	459	93.2
Clone 2 [†]	452	99.8		
Clone 3	333	97.3	303	98.0
Clone 4	400	98.5	*	*
Clone 5	454	98.0	454	96.3

* not available

† second sample preparation (sec. 4.2.3)

The second sequencing reaction performed with clone 2 gave the highest accuracy value, 99.8%. The malaria DNA is A and T rich. In the second sample prepared from clone 2 the labelling scheme was reversed so that the G and C terminated fragments were the most intense peaks. The greater number of lower

peaks from the abundant A and T terminated fragments did not overshadow the small G and C peaks, making interpretation of the sequence more reliable.

Slab gel sequence data was also obtained with the Pharmacia ALF sequencer for four of the five clones used in this study. This sequence data was also compared with the consensus sequence; the results are summarized in Table 4.1. The accuracies for the slab gel data ranged from 93.2% to 99.3%.

The average accuracy of the capillary gel sequence data, including errors and ambiguities, is 98.5%. This value compares favourably with the average accuracy of 96.7% for the Pharmacia ALF automated sequencer, determined experimentally for the same clones. Hood and coworkers (10) claim an average error plus ambiguity rate of 1 to 5% (95 to 99% accuracy) per run using the ABI 373A automated sequencer. The accuracy of the malaria sequence obtained by CGE could be increased by changing the labelling scheme to take advantage of the particular nature of the malaria DNA, as demonstrated by the data obtained for clone 2.

4.3.2 Sequence Length

The DNA sequences determined by CGE gave sequences of 330 to 460 bases in length (Table 4.1). The length of sequence obtained is comparable to that obtained by automated slab gel electrophoresis. One problem that affects both sequencing techniques is the quality of the DNA template that is to be sequenced. The amount of sequence data generated is dependent on the template. For example, clone 3 could only be sequenced to about 300 base pairs using slab and capillary techniques. Malaria is A/T rich, leading to fewer problems with compressions which are characteristic of a G/C rich sample. A typical run usually yields a sequence about 400 nucleotides long, and an exceptional run results in a sequence of 500 nucleotides or longer.

4.3.3 Sequencing Rate

Capillary gel electrophoresis runs faster than the slab gel technique. The sequencing rates varied from 230 to 450 bases/hour for the capillaries run at 200 V/cm, the slower rates resulting from longer capillary lengths. Total run time, at 200 V/cm, was on average 2.5 hours. The automated Pharmacia sequencer typically requires 6 to 8 hours to generate a sequence of about the same length. Slab gel is limited to the low field of approximately 50 V/cm due to the effects of Joule heating. The CGE data were taken at 200 V/cm, and achieved about a threefold decrease in runtime. For example, sample 1 required a 6 hour run on slab gel, whereas the same number of bases were completed in 1.5 hours with CGE.

4.3.4 Resolution

The capillary gel electrophoresis system operates at an electric field of 175 to 200 V/cm, which is roughly four times higher than the electric field used in the ALF instrument. According to classic electrophoresis theory, the resolution produced in an electrophoretic separation should scale with the square root of voltage (not electric field) used for the separation (15). The ALF system operates at an electric field of about 50 V/cm over a separation distance of about 40 cm; the separation voltage is about 2000 V. The capillary system is run at an electric field of 200 V/cm over a distance of 40 cm; the separation voltage is about 8000 V. The capillary electrophoresis system should have a resolution that is $\sqrt{8000 / 2000}$ or 2 times larger than the slab gel instrument, assuming mass diffusion dominates the separation.

Figure 4.6 presents the raw data obtained from the ALF instrument and the capillary gel electrophoresis instrument for fragments from 359 to 388 bases in length. The resolution of the ALF system is 0.5, measured for bases 360-361. The capillary gel data have a resolution of 1, measured for the doublet A bases 363-364. The two-fold improvement in resolution is exactly as predicted by classic electrophoresis theory. For fragments in this size range, capillaries produce higher resolution than slab gels because the capillaries operate at a higher voltage.

4.4 Conclusions

The accuracy of capillary gel electrophoresis is comparable to slab gel electrophoresis. The quality of the sequence data is template dependent for both techniques; clean, A/T rich DNA templates lead to the best sequencing runs. The agreement of sequence data between the two techniques is very high, greater than 97.5%. Capillary gel electrophoresis with the two-colour peak-height encoded sequencing technique produces superior resolution compared with commercial sequencing instruments. While not the primary purpose of this study, the CGE separation of DNA sequencing fragments is at least 3 times faster than slab gel electrophoresis. The main advantage of CGE is its ability to produce data of the same quality as automated slab gel electrophoresis in a shorter time. Clearly, a multiple capillary system coupled with an efficient algorithm to identify the DNA sequence would be significantly faster than the current automated systems.

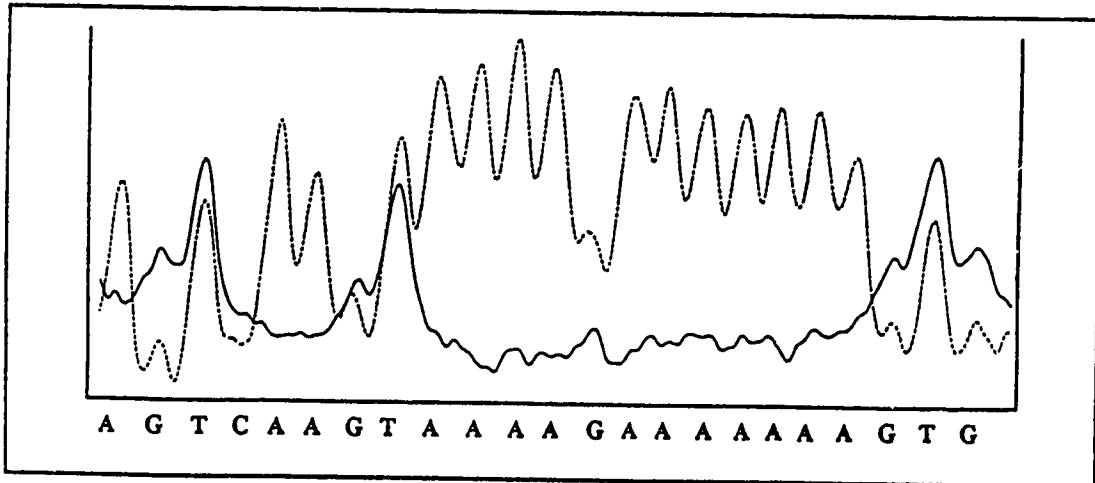
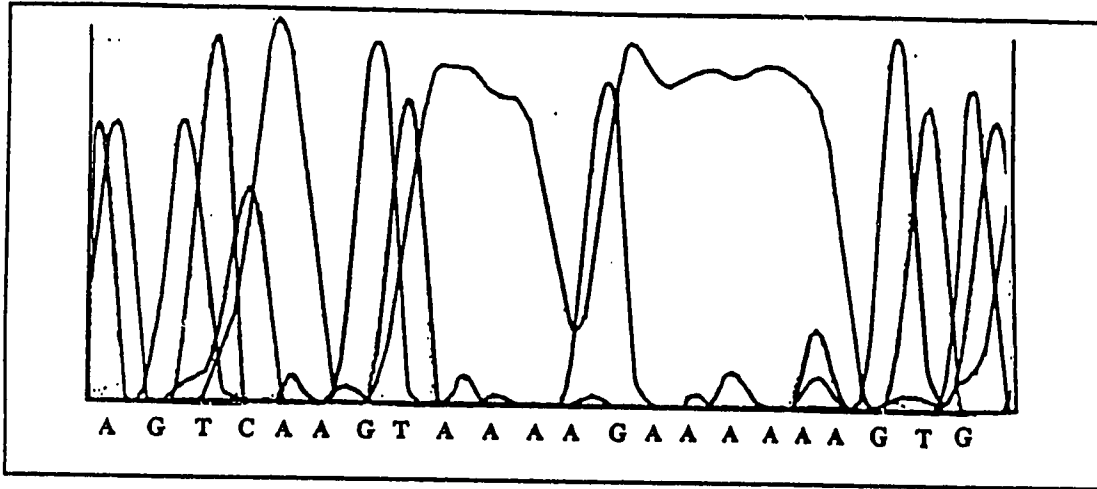


Figure 4.6 Comparison of resolution for fragment 359 to 388 bases in length. the top panel is data from the ALF sequencer. The bottom panel is data from the two-colour CGE instrument.

References

1. F. Sanger, S. Nicklen, A.R. Coulson, (1977), *Proc. Natl. Acad. Sci. USA*, **74**, 463.
2. L.M. Smith, J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connell, C. Heiner, S.B.H. Kent, L.E. Hood, (1986), *Nature*, **321**, 674.
3. W. Ansorge, B.S. Sproat, J. Stegemann, C Schwager, (1986), *J. Biochem. Biophys. Meth.*, **13**, 315.
4. J.M. Prober, G.L Trainor, R.J. Dam, F.W. Hobbs, C.W. Robertson, R.J. Zagursky, A.J. Cocuzza, M.A. Jensen, K. Bauermeister, (1987), *Science*, **238**, 336.
5. A.S. Cohen, D.R. Najarian, B.L Karger, (1990), *J. Chromatogr.*, **516**, 49.
6. H. Swerdlow, R. Gesteland, (1990), *Nucl. Acids Res.*, **18**, 483.
7. D.Y. Chen, H.P Swerdlow, H.R. Harke, J.Z. Zhang, N.J. Dovichi, (1991), *J. Chromatogr.*, **559**, 237.
8. L.G. Lee, C.R. Connell, S.L. Woo, R.D. Cheng, B.F. McArdle, C.W. Fuller, N.D. Halloran, R.K Wilson, (1990), *Nucl. Acids Res.*, **20**, 2471-2483.
9. G.L. Trainor, (1990), *Anal. Chem.*, **62**, 418-426.
10. T. Hunkapillar, R.J. Kaiser, B.F. Koop, L. Hood, (1991), *Science*, **254**, 59-67.
11. D.Y. Chen, H.R. Harke, N.J. Dovichi, (1992), *Nucl. Acids Res.*, **20**, 8322.
12. S. Tabor, C.C. Richardson, (1990), *J. Biol. Chem.*, **265**, 8322-8329.
13. S. Tabor, C.C. Richardson, (1989), *Proc. Natl. Acad. Sci. USA*, **86**, 4076-4080.
14. B.F. Koop, L. Rowan, W.Q. Chen, P. Deshpande, H. Lee, L. Hood, *BioTechniques*, (1993), **14**, 442-447.
15. J.W. Jorgenson, K.D. Lukacs, *Anal. Chem.*, **53** (1981), 1298.

CHAPTER 5

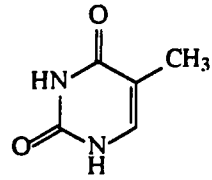
Internal Fluorescence Labelling with Two-Label Peak-Height Encoded DNA Sequencing by Capillary Gel Electrophoresis⁴

5.1 Introduction

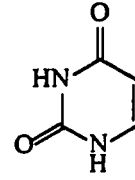
Current automated DNA sequencing technology makes use of fluorescent labels to detect and identify the DNA fragments. The labelled fragments are generated either by use of labelled primers or by incorporation of labelled dideoxynucleotide triphosphates (ddNTPs) (1-3). In 1992 Voss et al. reported the use of both fluorescein-12-dUTP (F-12-dUTP) and fluorescein-15-dATP (F-15-dATP) as internal labels for DNA sequencing with T7 polymerase and demonstrated this labelling strategy for primer walking applications (4-5). More recently, Hou and Smith demonstrated hexamer priming with internal fluorescence labelling (6).

The use of fluorescently tagged deoxynucleotides presents two advantages over the current labelling techniques. First, the incorporation of the fluorescent dNTP in a separate labelling step is compatible with a wide range of primers, hence time consuming and costly preparation of fluorescently labelled primers for primer walking is avoided. Second, the incorporation of these dNTP labels is quite uniform. Extensions and terminations carried out with T7 polymerase and manganese demonstrate uniform peak intensities, which increase the accuracy of the sequence obtained. Uniform terminations are not obtained, even under optimized experimental conditions with the labelled ddNTPs (3). Third, the available dye labelled

⁴A version of this chapter has been accepted for publication in Nucleic Acids Research.



Thymine (T)



Uracil (U)

Figure 5.1 The bases thymine and uracil.

deoxynucleotides are much less expensive than fluorescently labelled dideoxynucleotides or fluorescently labelled primers. Deoxyuridine triphosphate (dUTP) contains the base uracil (U) which is similar in structure to thymine (Figure 5.1), and the fluorescein tagged dUTP is readily incorporated by T7 polymerase in place of dTTP (4).

Chapter 3 describes the two-label peak-height encoded DNA sequencing technique (7-8). This method relies on the uniform incorporation of ddNTP's achieved by T7 polymerase in the presence of manganese (9). The resultant sample may be separated on a single lane of a slab gel or in a single gel filled capillary tube. Samples labelled with fluorescent primers demonstrate the uniform peak heights required for this method.

For the two-label peak-height encoded technique to be useful with capillary gel electrophoresis, two spectrally different labels are required. Until recently, only fluorescein labelled dNTPs have been commercially available, limiting their use in internal labelling with peak-height encoding. Boeringer Mannheim kindly supplied a tetramethylrhodamine labelled dATP (TR-dATP). Simple modification of the previously reported detection system allows detection of both the fluorescein and the tetramethylrhodamine labels. This chapter describes the use of internal labelling with F-12-dUTP and the two-label peak-height encoded sequencing technique using labelled dATPs. The separation of these internally labelled samples are compared with the separation of primer labelled samples.

5.2 Experimental

5.2.1 Instrumental Design

The system used for DNA sequencing by capillary electrophoresis is similar to those previously reported (7,10). For the F-12-dUTP labelled samples the capillaries are 50 μm i.d. and 190 μm o.d, and 40 cm long. The capillaries were filled with 5%T 6%C polyacrylamide gel as described in Chapter 3. The instrument is described in detail in Sec. 2.2.1. For separation of the fluorescent dATP labelled samples the capillary tubing was 50 μm i.d., 145 μm o.d., and typically 35 cm long. The capillaries were filled with non-crosslinked polyacrylamide. The polyacrylamide was prepared from 5 ml aliquots of acrylamide monomer solution (6%T), 1X TBE, and 7M urea. The monomer solution was degassed by bubbling helium through the solution for a period of 5 minutes. Polymerization was then initiated by addition of 2 μl of TEMED and 20 μl of 10% ammonium persulphate. The capillaries were treated with a solution of γ -methacryloxypropyltrimethoxysilane to bind the polyacrylamide to the capillary walls. The capillaries were filled with the monomer solution by application of a vacuum. Polymerization appears complete within 30 min., however, capillaries were typically stored overnight before use.

The injection end of the capillary was held inside a Plexiglas box equipped with a safety interlock. The other end of the capillary was inserted into the flow chamber of a locally constructed sheath flow cuvette. The cuvette had a 150 μm square flow chamber and 1 mm thick quartz windows.

5.2.2 Detector

The fluorescence detector is shown in figure 5.2. A 10 mW argon ion laser beam ($\lambda=488\text{nm}$) is aligned with a dichroic filter to be coincident with a 2 mW helium-neon laser beam ($\lambda=543.5\text{nm}$). The combined laser beams are focussed with a 5X microscope objective onto the sheath flow cuvette about 100 μm below the tip of the capillary. Fluorescence is collected at a right angle to the laser beams with a 0.6 NA, 32X microscope objective (Leitz/Wild model 2569-1130). The fluorescence is imaged onto a 0.75 mm diameter pinhole. A 545 nm dichroic filter splits the fluorescence into two spectral channels. The transmitted fluorescence passes through a spectral bandpass filter centered at 580 nm with a 30 nm bandwidth. The reflected fluorescence passes through a spectral bandpass filter centered at 515 nm with a 25 nm bandwidth. In both channels the fluorescence is collected with an R1477 photomultiplier tube (PMT) (Hamamatsu). The output from each PMT is conditioned with a low pass electronic filter with a time constant of 0.5 sec, and digitised by a National Instruments A/D board in a Macintosh IIsx computer. Data is collected at a sampling rate of 2 Hz. Before presentation, the data is processed to compensate for a mobility shift and is smoothed with 2 passes of a binomial filter. The sheath flow is 1X TBE provided by a simple siphon based on a 4 cm height difference between the sheath buffer reservoir and the waste collection vial. The sheath flow rate is approximately 0.08 ml/hr.

5.2.3 Sample Preparation

Fluorescein-12-dUTP labelled samples were prepared by simple modification of the sequencing protocol outlined in sec. 3.2.3. Unlabelled M13 -21 universal primer

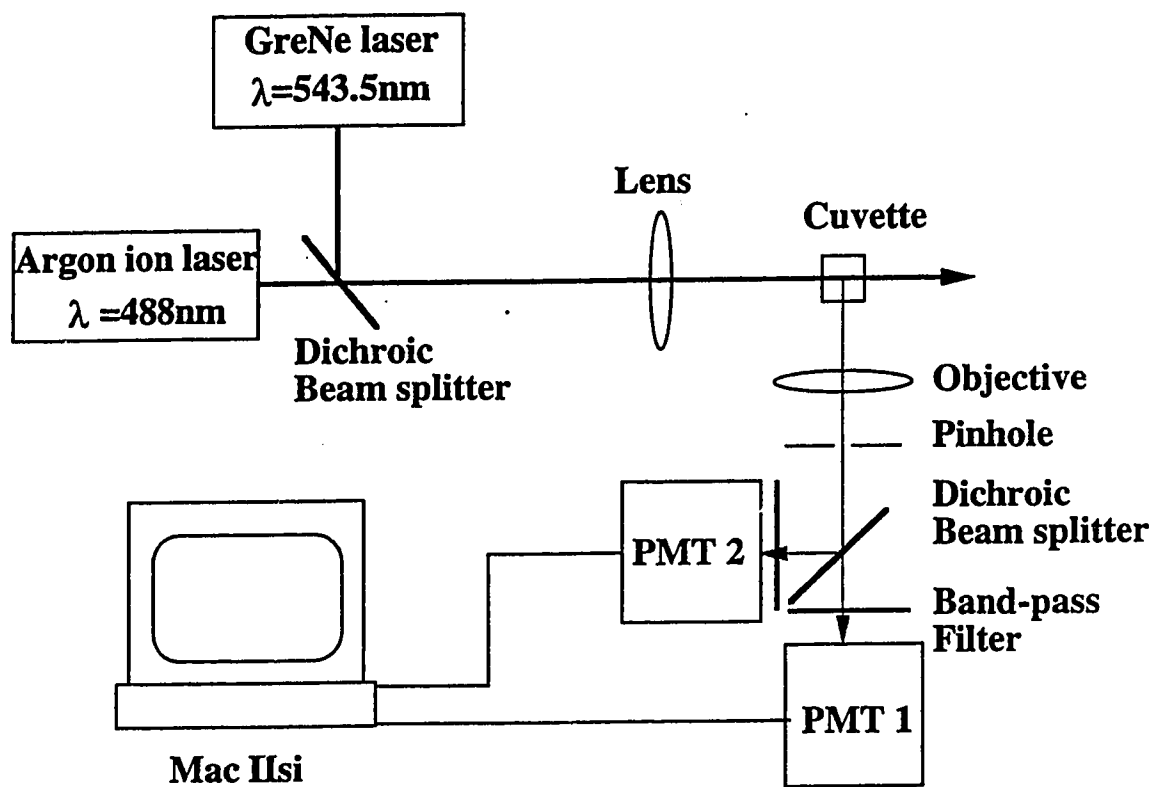


Figure 5.2 Diagram of the two-laser, two-channel fluorescence detector.

was annealed to M13mp18 single stranded template DNA. The reaction mixture consisted of 2.5 μ l of 10X MOPS buffer (400 mM MOPS pH 7.5, 500 mM NaCl, 100 mM $MgCl_2$), 2.5 μ l 10X Mn solution (50 mM $MnCl_2$, 150 mM sodium isocitrate), 1 μ g M13mp18 single stranded DNA (0.2 mg/ml) (USB, Cleveland, OH), 2 μ l M13 -21 primer (0.5 mM) (USB) and water to give 18 μ l. To anneal the primer to the template the mixture was heated to 65°C for two minutes, followed by slow cooling to room temperature. The labelling step was accomplished by addition of 2 μ l of labelling mix (1.5 μ M each dATP, dCTP, dGTP, and F-12-dUTP) and 6 units of Sequenase (USB). The mixture was incubated at 18°C for 10 minutes. Following the labelling step, 10 μ l of the combined A and C termination mix was added (A and C termination mixes were mixed in a 3:1 ratio A:C; each mix was 1 mM each dATP, dCTP, dGTP, dTTP, and 3.3 μ M of the appropriate ddNTP). The mixture was incubated at 37°C for 10 minutes to allow the chain extension and termination reactions to proceed. The reaction was stopped by addition of 12 μ l stop/salt solution (20 mM EDTA, 1 M sodium acetate, pH 8.0) and the DNA was precipitated by addition of 120 μ L 98% ethanol. The solution was kept at -20°C for at least 30 minutes, typically overnight. Following centrifugation and washing with 300 μ L ice cold 80% ethanol the DNA was resuspended in 4 μ l formamide-0.5 M EDTA (49:1).

The two-colour sequencing reactions were carried out in buffers from the Pharmacia AutoRead sequencing kit. The G and T terminations were prepared by mixing 1 μ g of M13mp18 single stranded DNA with 2 μ l of -40 primer (5 μ M), 2 μ l of annealing buffer (1M Tris-HCl, pH 7.4, 100mM $MgCl_2$), and H_2O to a volume of 15 μ l. The primer and template were annealed as above. The labelling reaction was carried out by addition of 1 μ l of extension mix (40mM $MnCl_2$, 302 mM sodium citrate, 324 mM dithiothreitol (DTT)), 2 μ l labelling mix (10 μ M F-15-dATP, 1 μ M

each dCTP, dGTP, and dTTP), and 6 units of Sequenase (USB). The reaction was incubated at 37°C for 10 min. The labelling reaction was followed by addition of 3.5 µl of DMSO and 10 µl of the combined G and T termination mixes (mixed in a 3:1 ratio G:T; each mix is 1 mM in each dATP,dCTP, dGTP,dTTP and 5 µM in the respective ddNTP, 50 mM NaCl, 40 mM Tris-HCl, pH 7.4). The termination reaction was carried out at 37°C for 10 min. The reactions were stopped by addition of 13 µL of stop solution (1 M NaOAc, pH 8.0, 20 mM EDTA) and the DNA was precipitated. Identical conditions were used to generate the tetramethylrhodamine labelled sample. The labelling mix contained 10µM TR-dATP and the A and C termination mixes were added in a 3:1 ratio. The samples were each resuspended in 3 µl of a 49:1 mixture of formamide-0.5 M EDTA. The F-15-dATP labelled sample was mixed with the TR-dATP labelled sample.

The MMTV sample was prepared as above by annealing 1 µg of MMTV template DNA with 2 µl of unlabelled M13 -21 primer (0.5 µM). The malaria sample was prepared with 1 µg of malaria template DNA with 1 µl of unlabelled T3 primer (1 µM).

To study the products of the labelling reaction an F-15-dATP labelled sample was prepared. Unlabelled M13 -21 primer and M13mp18 template DNA were annealed and the labelling reaction was carried out as outlined above. The labelling reaction was followed by addition of stop/salt solution to stop the reaction and the DNA was precipitated with ethanol. The sample was resuspended in 4 µl of formamide-0.5 M EDTA. One microlitre of TAMRA primer (20 nM) was added to the sample as an internal size standard; the primer is a 17mer.

For the comparison study, ddC terminated M13mp18 samples were prepared in the manner described above. The internally labelled sample was labelled with F-15-dATP.

The primer labelled sample was labelled with a locally synthesized, FAM labelled, -40 M13 primer, and the labelling step was eliminated from this synthesis.

5.2.4 Electrophoresis

The F-12-dUTP labelled sample was separated on a 4%T 5%C polyacrylamide gel. The capillary was 42 cm long. The sample was heated to 95°C for two minutes and then injected onto the capillary by applying a 200 V/cm electric field for 30 sec. Electrophoresis was carried out at 200 V/cm.

The two-colour dATP labelled sample was separated on 6%T linear polyacrylamide. The capillary length was 36 cm. The sample was injected for 60 sec at 200 V/cm and electrophoresis was carried out at 200 V/cm.

The products of the labelling reaction were separated on a 6%T linear polyacrylamide filled capillary, 37 cm long. The sample was injected for 15 sec at 200 V/cm and electrophoresis was carried out at 200 V/cm.

For the comparison study, the samples were separated on 6%T 0%C filled capillaries, 35 cm long. The samples were injected for 60 sec. at 200 V/cm and electrophoresis was carried out at 300 V/cm.

5.3 Results and Discussion

5.3.1 F-dUTP Labelled Sample

A typical electropherogram is shown in figure 5.3, only the transmitted channel is plotted. The detector used is described in detail in sec. 3.2.1. The fluorescein label on the DNA fragments gives signal in both the reflected and transmitted channels. The

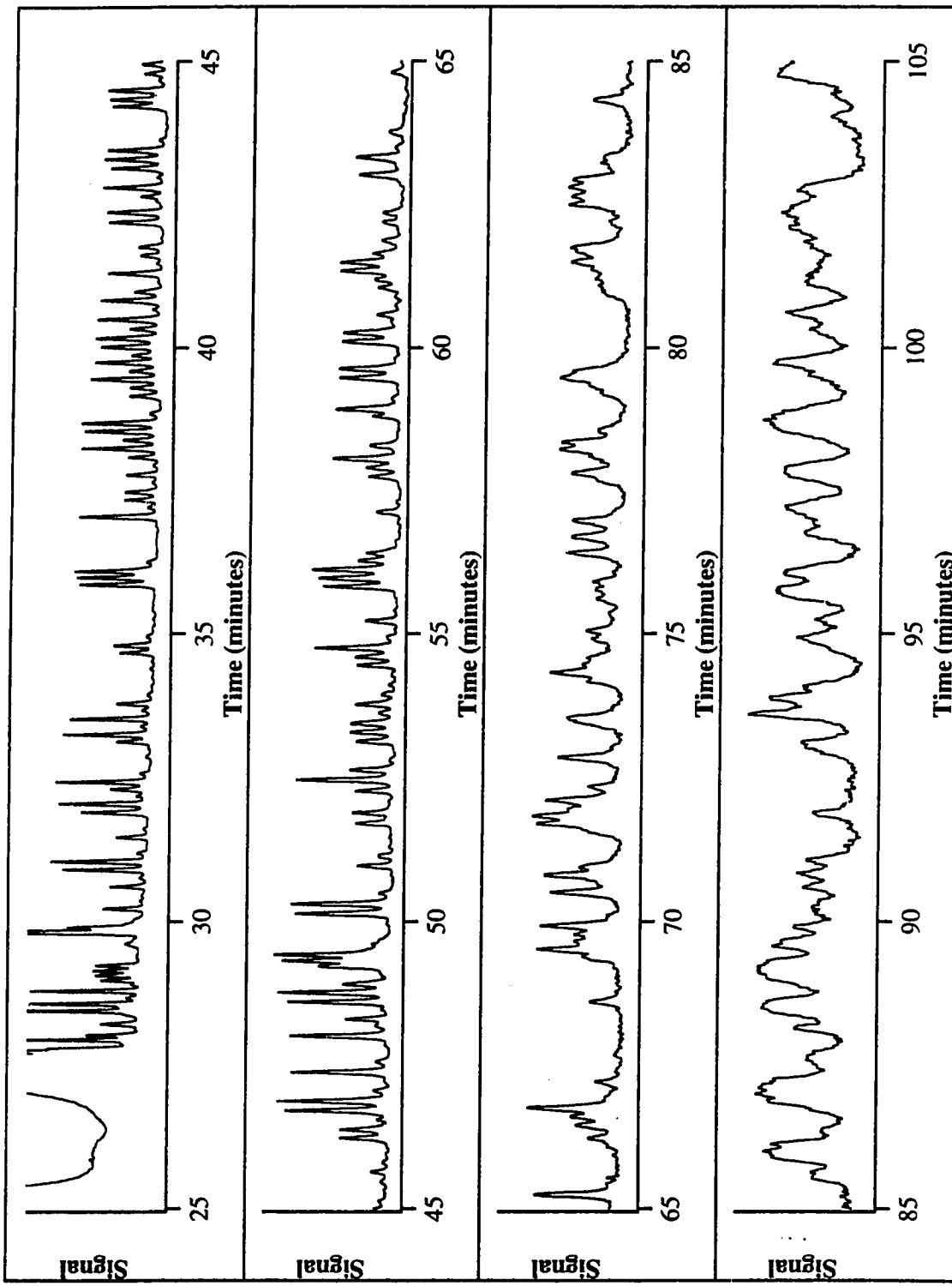
large peaks correspond to A's; the small peaks correspond to C's. The pattern of peaks in the electropherogram was correlated to the known sequence of M13mp18 for fragments 44 to 452 bases in length. One peak corresponding to a C at position 399 was missing. Some small ghost peaks are observed in the electropherogram, particularly in the early part of the separation. These are likely due to nonspecific priming and could be reduced or eliminated by reducing the amount of primer added to the sequencing reaction. In a two colour system, such as that described in chapter 3 for detection of FAM and JOE labels, ghost peaks can cause errors in identification of the small amplitude peaks. If, however, the spectral resolution of the two channels is improved such low intensity ghost peaks will not interfere with the sequence identification.

The F-12-dUTP is efficiently incorporated into the DNA strand by the polymerase; the signal to noise ratio is good and the uniformity of peak intensities and hence the peak height encoding of sequence is also good. Although only one fluorescently labelled dUTP is commercially available, this labelling system demonstrates that internal fluorescence labelling with peak height encoding is viable.

5.3.2 Two-Colour dATP Labelled Sample

The labelled dATPs are less efficiently incorporated by the T7 polymerase than the labelled dUTP. The concentration of dATP in the labelling mix is tenfold higher than dUTP, the labelling reaction is carried out at higher temperature, and the signal-to-noise ratio of the products is lower. In spite of the lower efficiency of incorporation and lower signal, the fluorescein labelled dATP is reported to be more

Figure 5.3 Electropherogram of an A and C terminated M13mp18 sample labelled with F-12-dUTP. The large peaks are A's and the small peaks are C's.



stable during storage and in the sequencing reactions than the fluorescein labelled dUTP (11).

Two lasers are used in the detector described in Sec. 5.2.2.

Tetramethylrhodamine is excited by the 5 mW green helium-neon laser operating at 543.5 nm and, to a lesser degree, by the argon ion laser operating at 488 nm.

Fluorescein is excited by the argon ion laser, operated 10 mW. Operating the argon ion laser at a relatively low power of 10 mW is a compromise. Higher laser power would give increased signal from the tetramethylrhodamine label; the fluorescein is photobleached at 10 mW incident laser power, so increasing the laser power provides no increase in fluorescein signal. Higher laser power also increases the background in the optical channel centered at 580 nm that detects the tetramethylrhodamine signal. Ideally, a higher power helium neon laser would be the best method of increasing the tetramethylrhodamine signal. The 5 mW laser used in this study is the highest power available at low cost.

The sequencing electropherogram is shown in figure 5.4. The bottom trace (transmitted channel) shows the A and C terminations; the large amplitude peaks are A, the small are C. The top trace (reflected channel) shows the G and T terminations; large peaks are G, small are T. The data was processed as described below. The two traces are offset for presentation. There is a small amount of cross-talk in the transmitted channel; some fluorescence from the F-15-dATP is transmitted by the 580 nm bandpass filter. A mobility shift between the two channels was also observed; the tetramethylrhodamine labelled fragments, detected in the transmitted channel, migrated faster than the fluorescein labelled fragments, detected in the reflected channel. This shift is presumably due to differences in the mobilities of the labels. The mobility shift was corrected by normalizing the two channels, subtracting the small

contribution from optical cross-talk in the transmitted channel, and shifting the transmitted channel data back by 10 seconds.

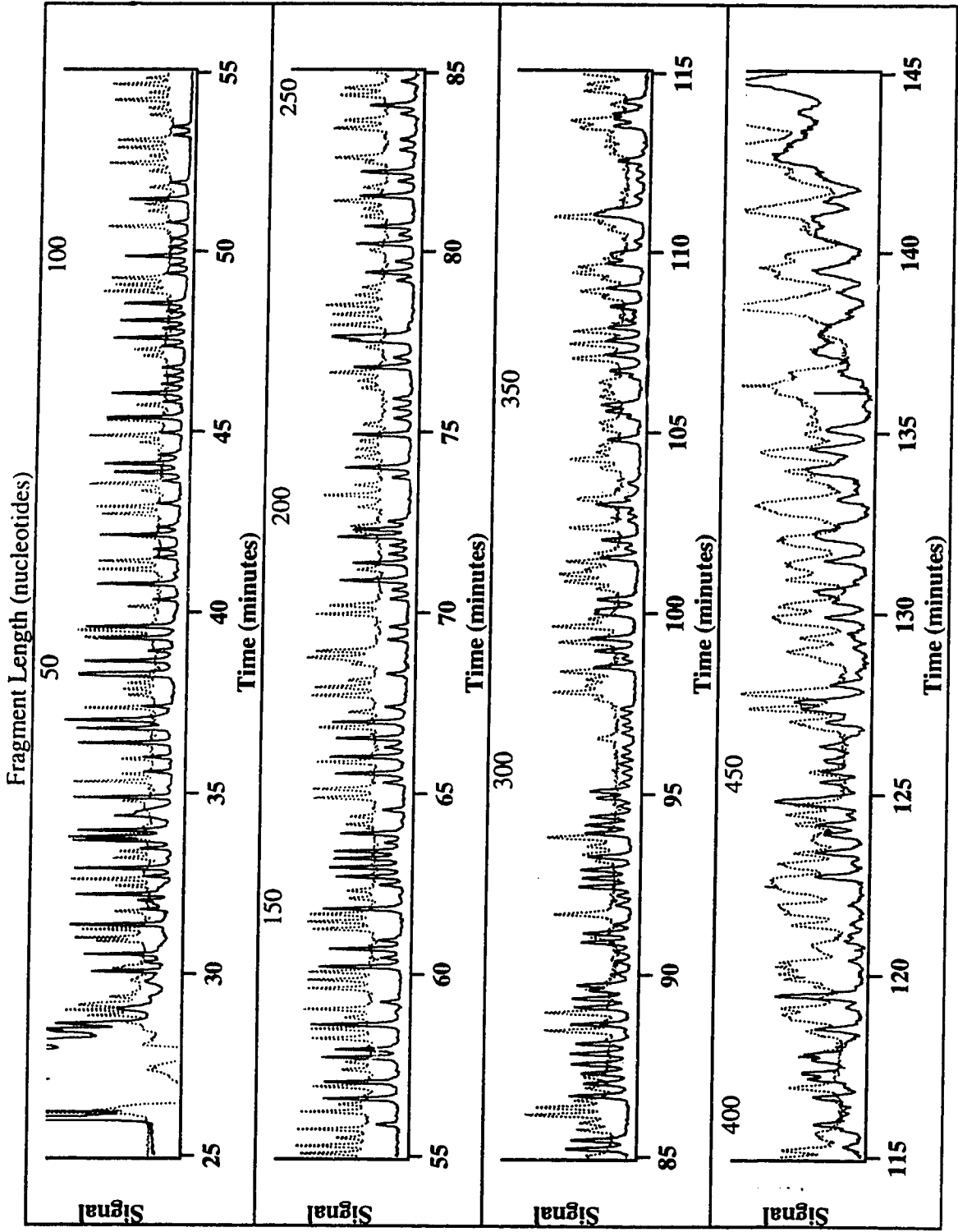
The DNA sequence may be identified for fragments 50 to 500 bases long. If the data are not processed to compensate for the mobility shift, peaks corresponding to C's (small peaks in the top channel) are frequently missed when they immediately follow a G (large peaks in the slower bottom channel). This effect is illustrated in figure 5.5. The top panel shows the smoothed data. The middle panel shows the data after correction for spectral cross-talk; the dashed trace was replaced by the original value minus 0.25 times the solid trace. The bottom panel shows the data after correction for mobility shift. The solid channel was offset by 7 seconds.

Figures 5.6 and 5.7 show internally labelled samples prepared from MMTV and malaria template after the data is processed in the manner described above. The signal to noise is slightly lower in these electropherograms. The sequence is identified to 400 bases for MMTV and to 320 bases for the malaria clone.

5.3.3 The Labelling Reaction

To further investigate the extent of incorporation of the labelled nucleotide, a reaction was stopped after the labelling step. The electropherogram is shown in figure 5.8. Depending on the experimental conditions, the primer is extended 6 to 8 nucleotides to a quartet of As. The main product (80 to 90 %) of this labelling reaction incorporates one labelled dATP. Minor products, incorporating two or three labelled dATPs, are observed. Based on these results, an investigation was undertaken to determine the effect of the internal labelling on the efficiency of the electrophoretic separation. The incorporation of more than one labelled nucleotide in fragments of a given length

Figure 5.4 Sequencing Electropherogram for the two-colour internally labelled M13mp18 sample.



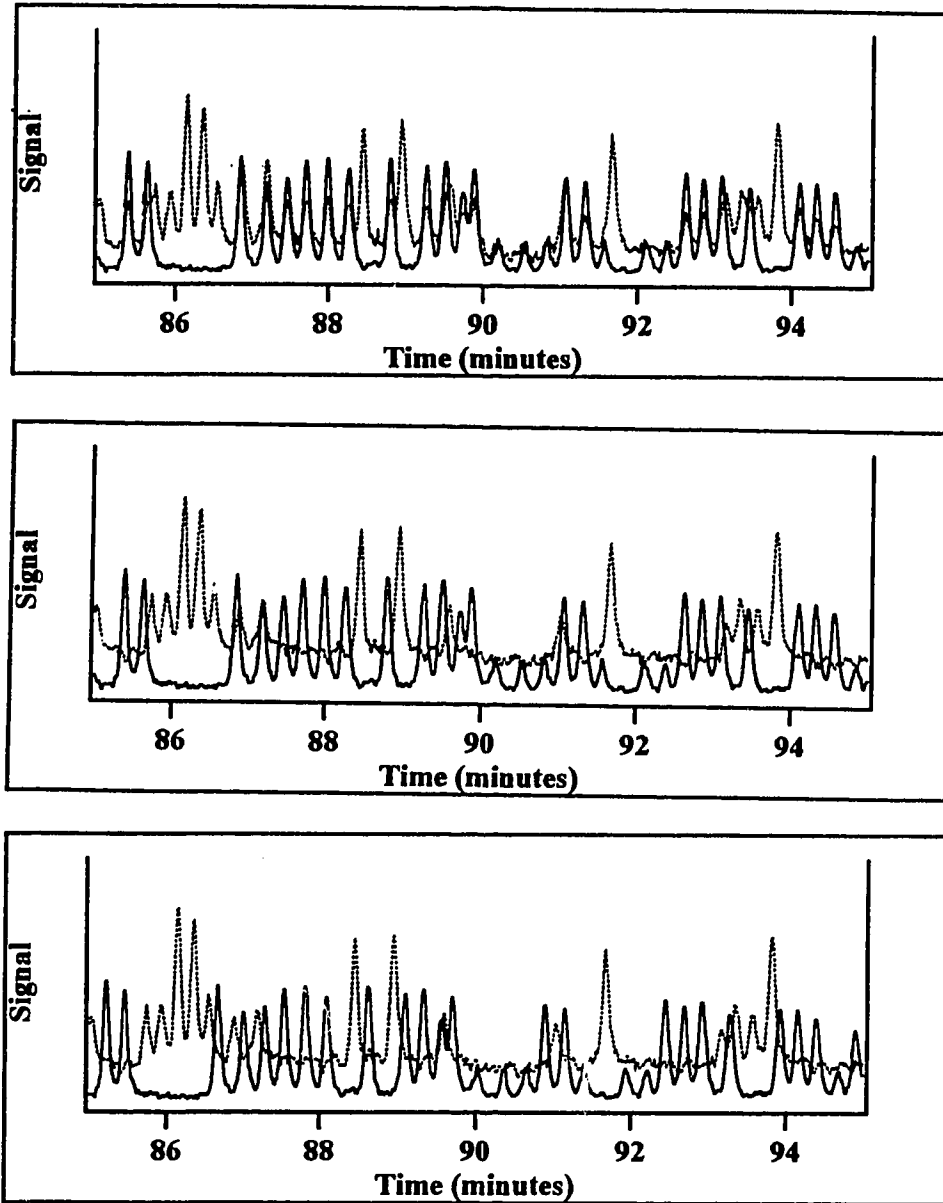


Figure 5.5 Data processing to compensate for the mobility shift. The top panel shows the smoothed data. The middle panel shows the data after correction for spectral cross-talk. The bottom panel shows the data after correction for the mobility shift.

Figure 5.6 Electropherogram of the internally labelled MMTV sample.

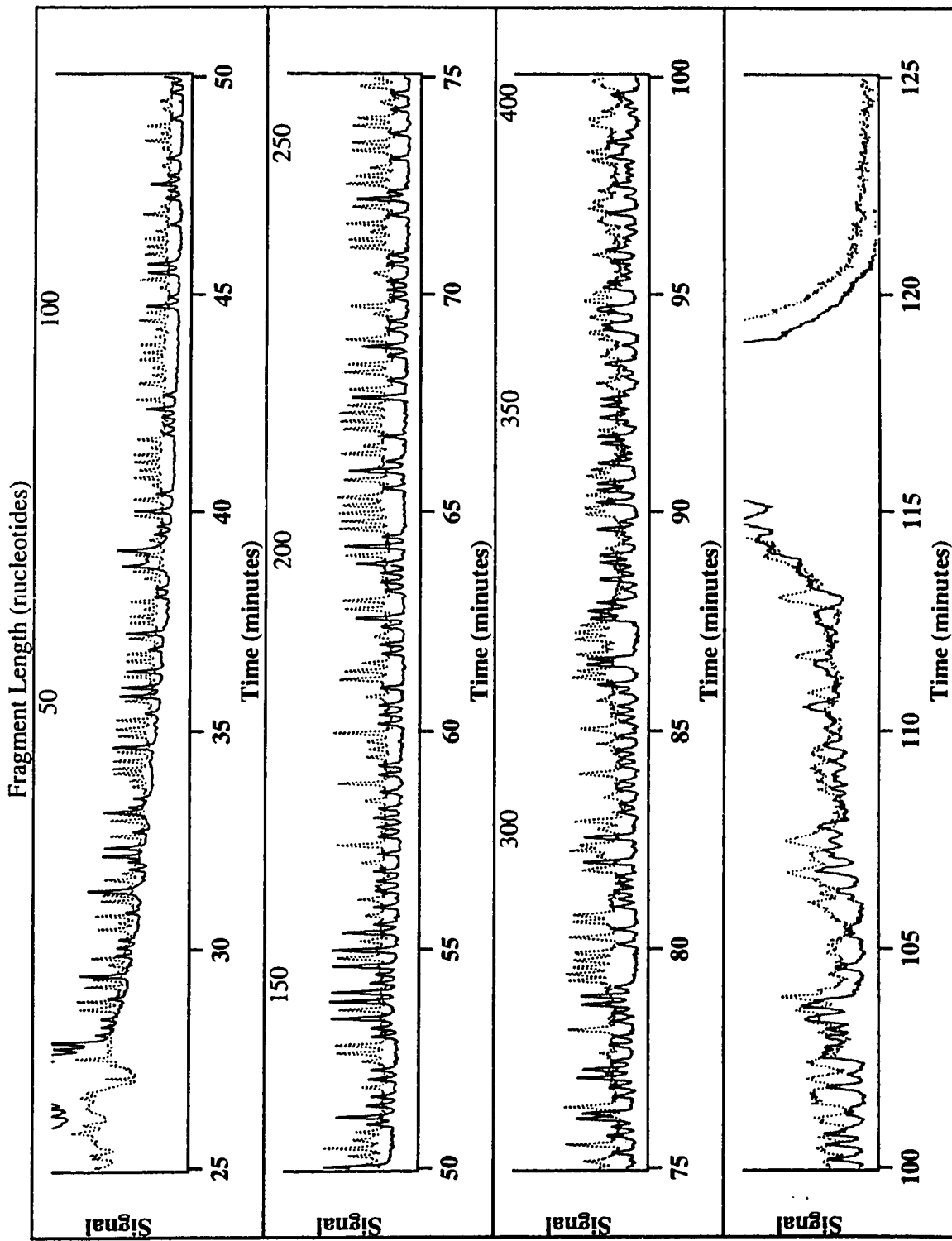
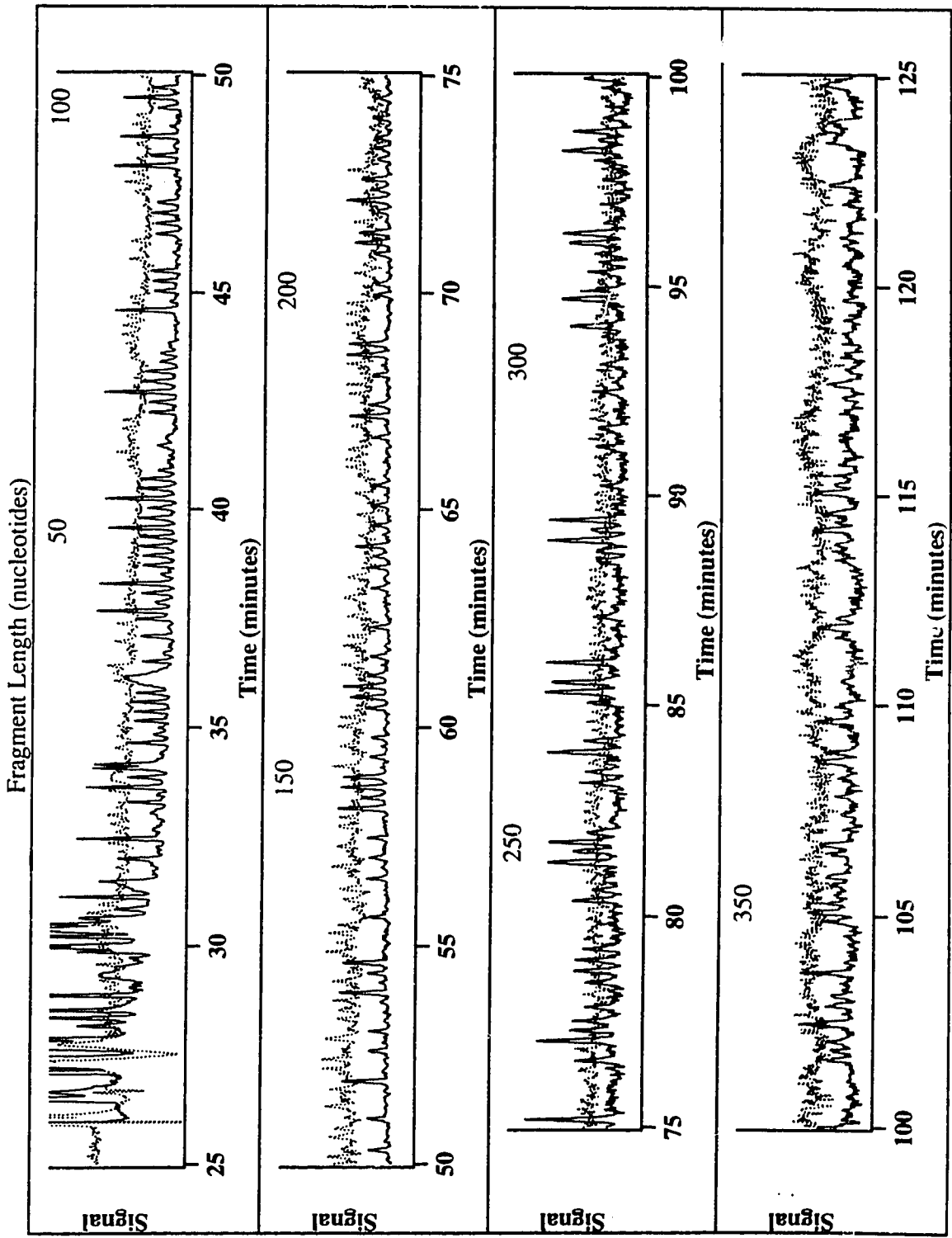


Figure 5.7 Electropherogram of the internally labelled malaria sample.



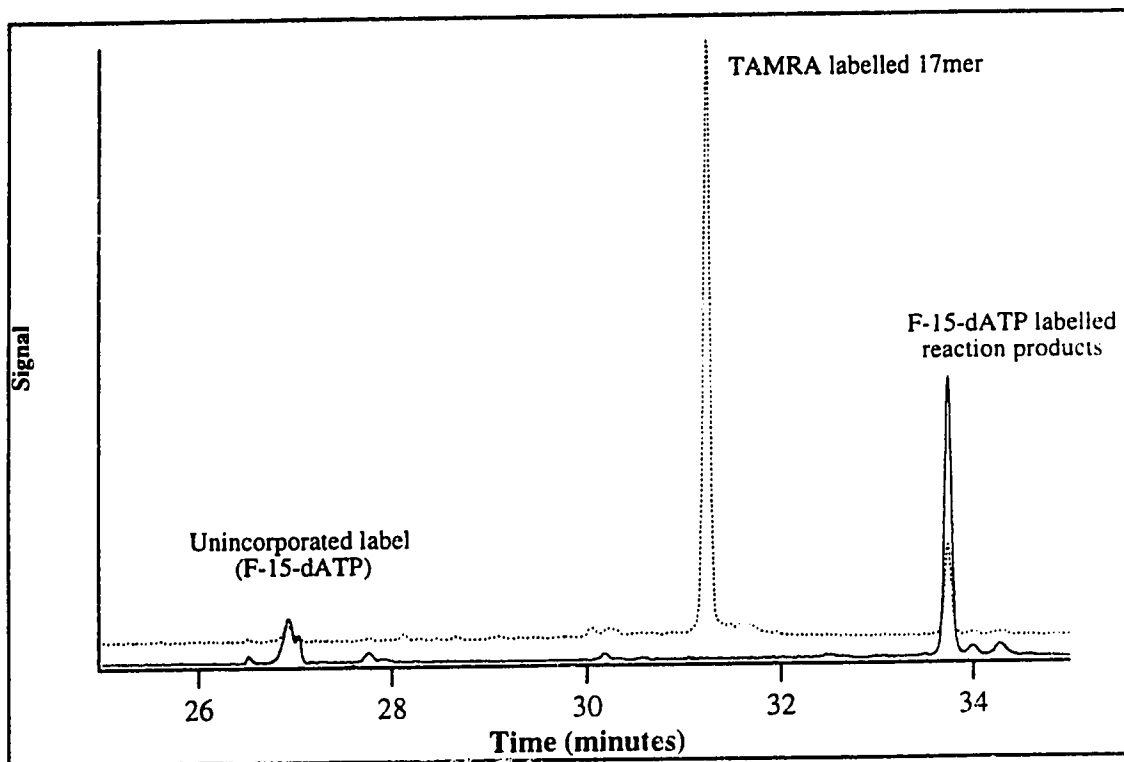


Figure 5.8 Electropherogram of the products of the labelling reaction. The solid trace shows the F-15-dATP labelled DNA fragments. The peak at 27 min is unincorporated label. Three peaks at 34 min are the products of the labelling reaction. the dashed trace shows a TAMRA labelled 17mer primer used as a size reference.

could cause the fragments to have a larger distribution of mobilities, and introduce peak broadening.

5.3.4 Comparison of Separation of Internal and Primer Labelled Samples

Two samples, one primer labelled, the other internally labelled, were run on separate, identical capillary gels. The gels were prepared from the same batch of acrylamide and were stored for the same length of time before the runs. During the experiment the lengths of the capillaries were the same within a few millimeters, and the prerun times were the same to within 3 minutes. The two runs were carried out in series, over a six hour period. This should present negligible differences in the gel composition inside the capillaries, and so the two runs took place under essentially identical conditions, allowing direct comparison of the results. The electropherograms are shown in Figure 5.9.

A plot of the retention times versus the fragment length is shown in Figure 5.10. The internally labelled fragments have longer retention times, hence lower mobilities than the primer labelled fragments of the same length. According to the biased reptation model, mobility is inversely proportional to fragment length

$$\mu = \chi \left[\frac{1}{N} + \frac{1}{N^*} \right] \quad (5-1)$$

where χ is a proportionality constant, N is the fragment length in bases, and N^* is the fragment length for the onset of biased reptation. Figure 5.11a presents plots of the inverse fragment length vs. mobility. A straight line was fit to the data for fragments longer than 200 bases in figure 5.11b; the onset of biased reptation for primer

Figure 5.9 a. Electropherogram of the internally labelled C terminated M13mp18 sample.

a.

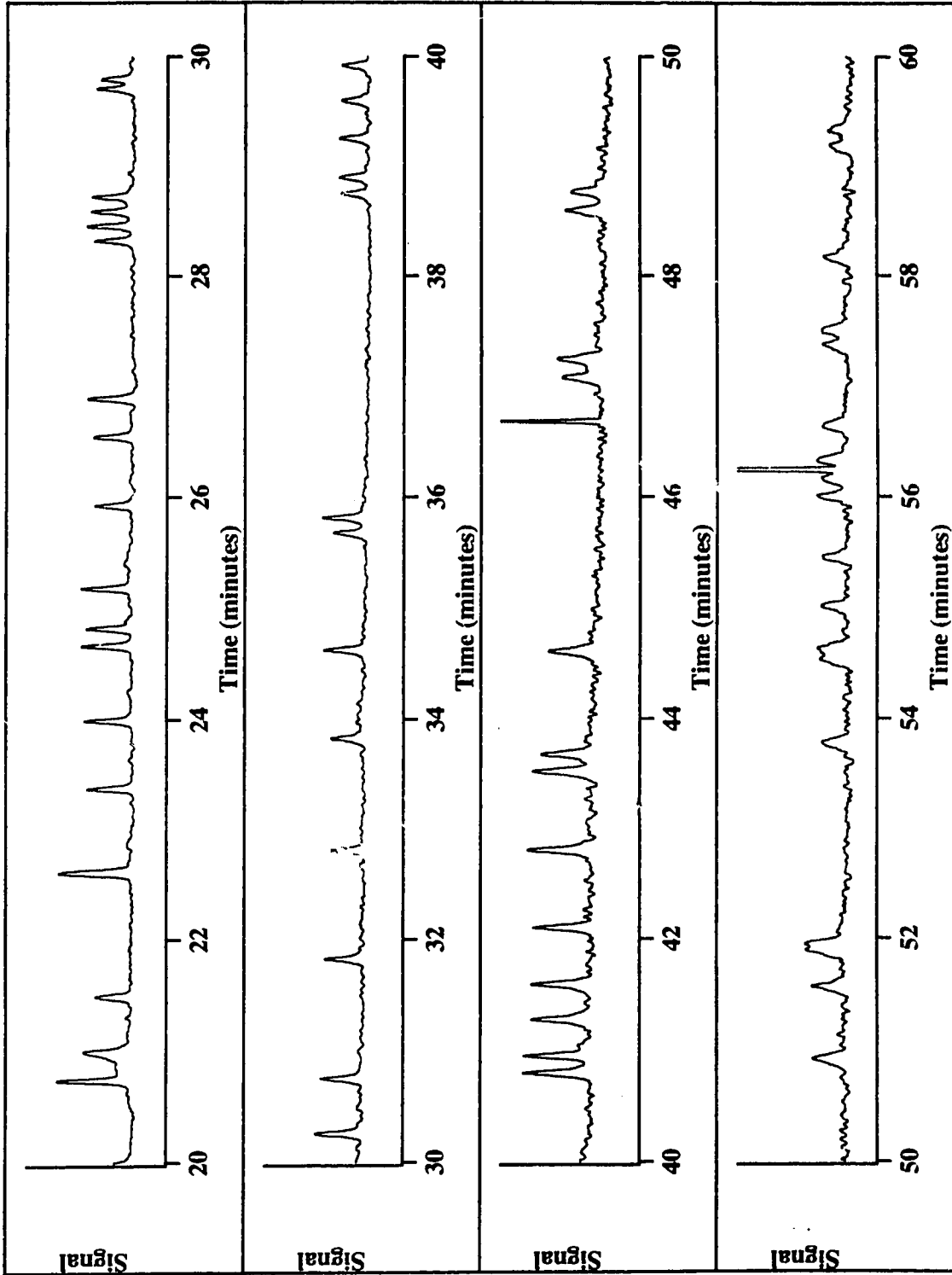
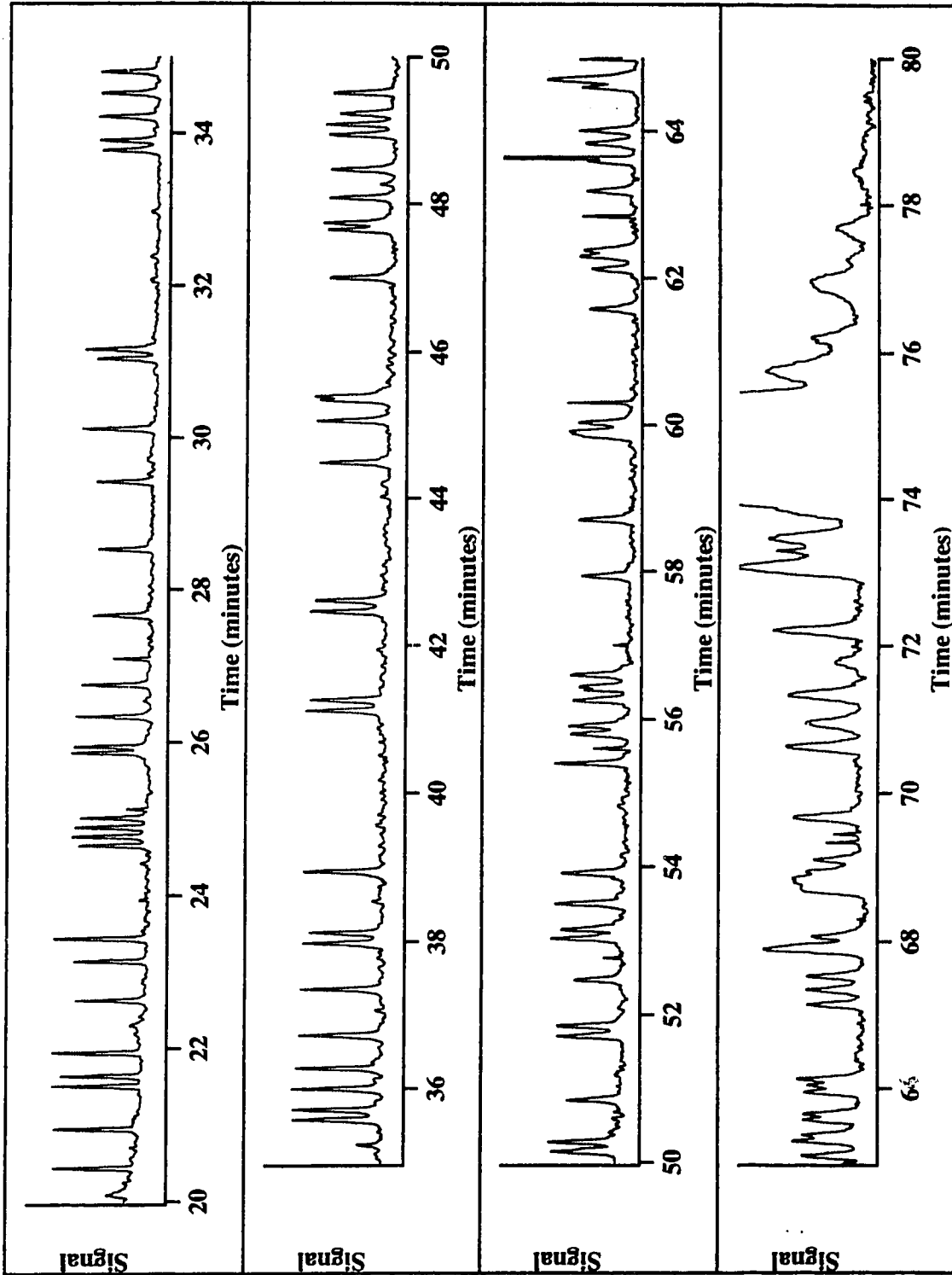


Figure 5.9 **b. Electropherogram of the primer labelled C terminated M13mp18 sample.**

b.



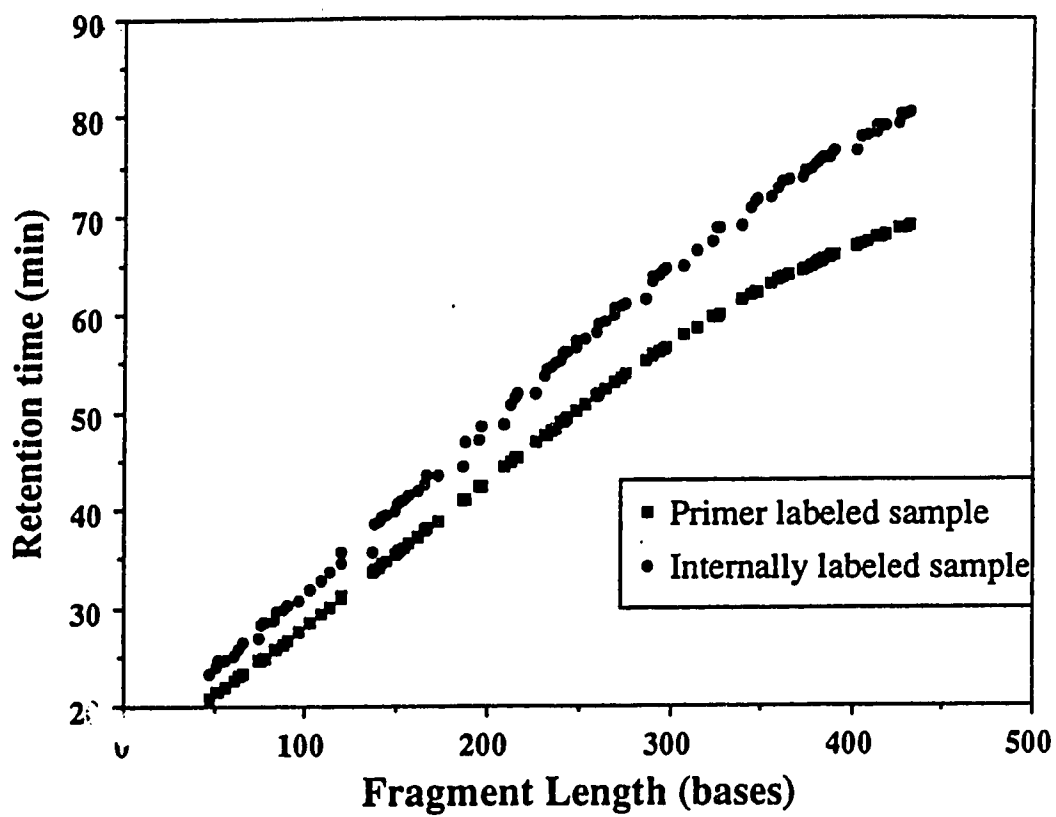


Figure 9.10 Retention time vs. fragment length for internal and primer labelled samples.

labelled fragments was observed at 490 bases while the internal labelled sample demonstrated an onset of biased reptation at 545 bases. The delayed onset of biased reptation for the internally labelled samples is likely responsible for the extremely long runs at low electric field reported by Ansorge and coworkers (5).

Resolution is used to describe the relative separation of adjacent peaks

$$R = \frac{\Delta T}{\overline{W}_{\text{baseline}}} \quad (5-2)$$

where R is resolution, ΔT is peak spacing, and $\overline{W}_{\text{baseline}}$ is the peak-width measured at the baseline. The resolution of adjacent peaks of all multiplets in the electropherograms was measured. Figure 5.12 presents resolution as a function of fragment length. Several points can be made about this resolution data. First, resolution decreases steadily with fragment length. Second, a compression is observed for fragments about 85 bases long. This compression results in an anomalously low resolution for both labelling schemes. Third, there appears to be a constant 0.4 difference in resolution between the two labelling schemes for fragments up to 300 bases in length. This consistently poorer resolution observed for the internally labelled fragments presumably is caused by a distribution of the number of labels incorporated in the sequencing fragments. Since the fluorescent label contributes to the mobility of the fragment, a distribution of labels will lead to a distribution of mobilities for fragments of identical length, leading to band broadening and degraded resolution. Fourth, similar resolution is obtained for longer fragments with both labelling techniques. Here, the onset of biased reptation causes fragments with greater than 350 bases to crowd together to a larger extent for primer labelled fragments compared with internally labelled fragments.

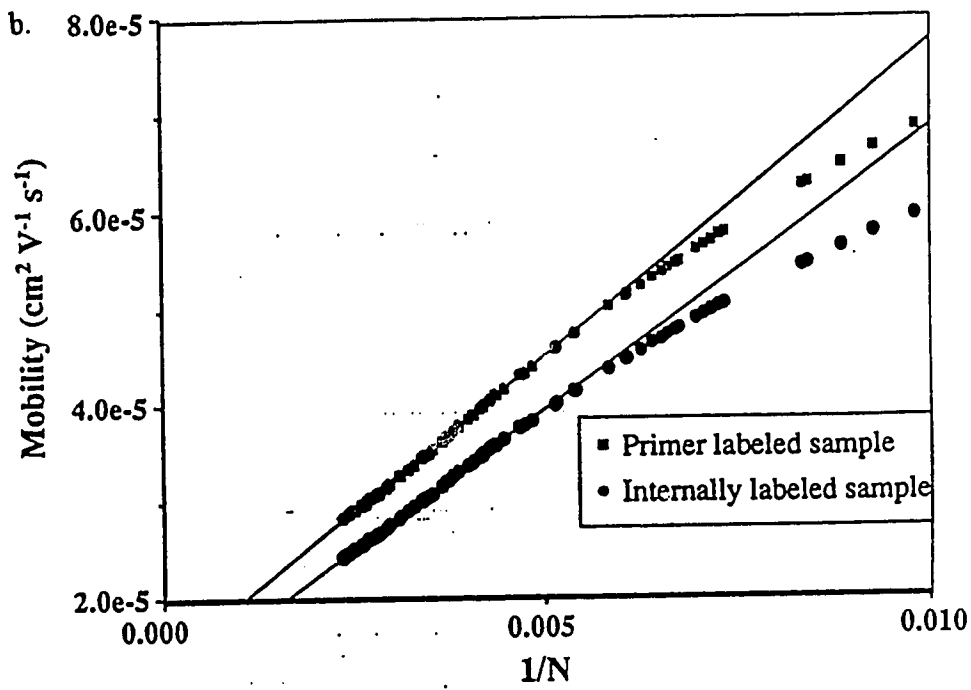
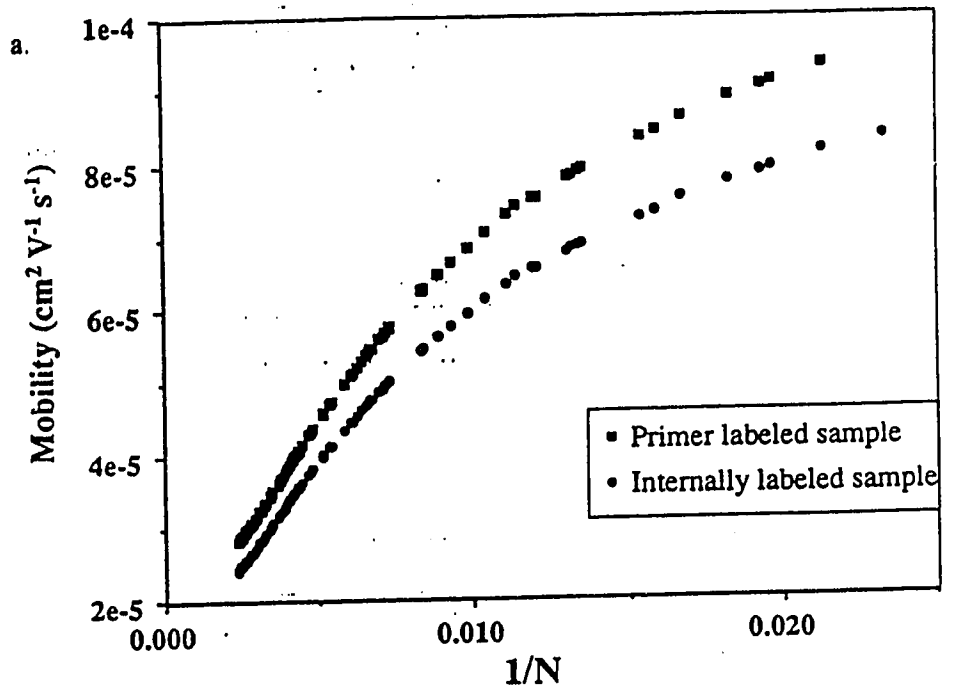


Figure 5.11 a. Inverse fragment length vs. mobility.
 b. A straight line fit to the data from fragments longer than 200 bases.

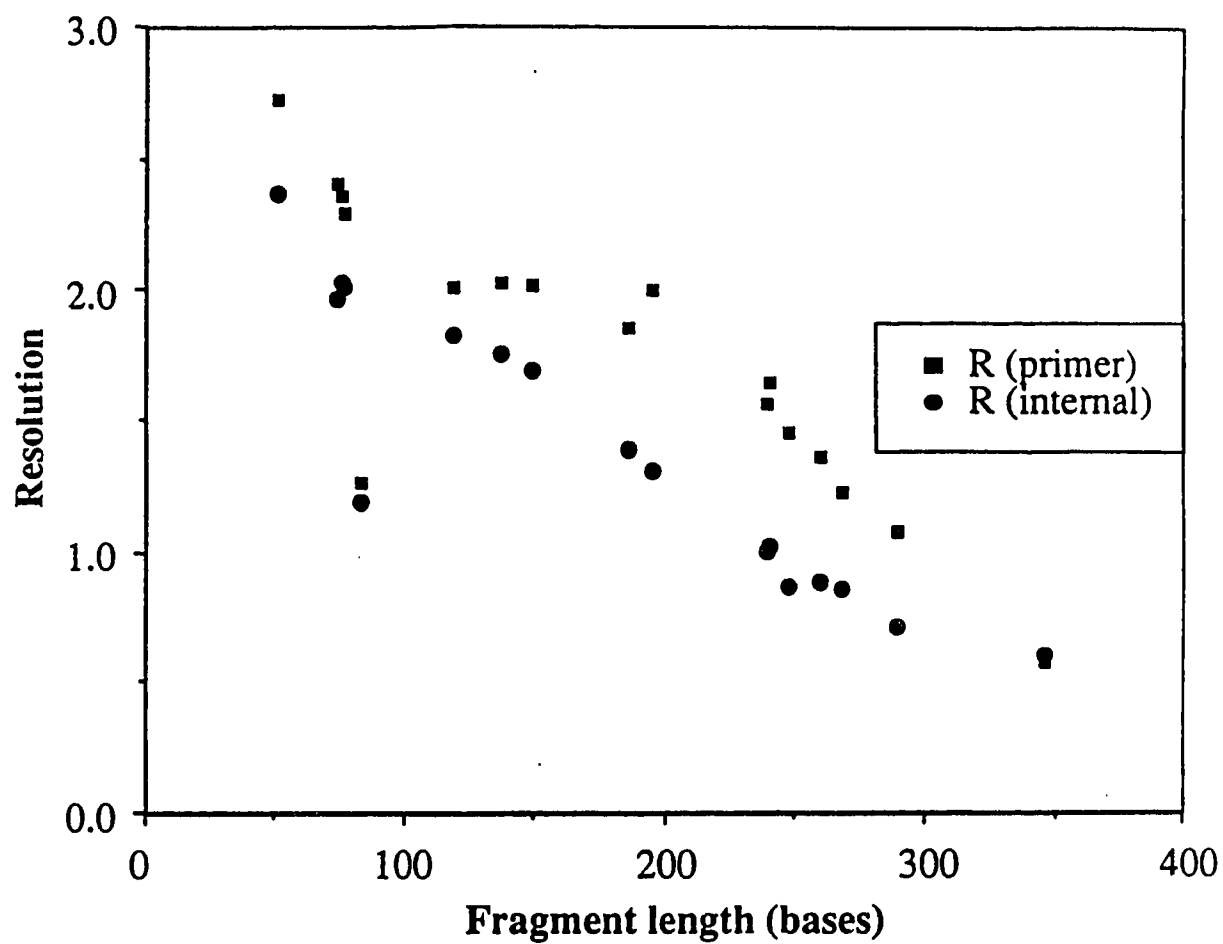


Figure 5.12 Resolution as a function of fragment length for primer and internal labelled samples.

5.4 Conclusions

The incorporation of the fluorescently labelled deoxynucleotides into DNA sequencing samples for the two-label peak height encoded technique is quite readily achieved.

The fluorescein-12-dUTP is more efficiently incorporated by the polymerase and gives higher signal-to-noise ratio than the fluorescein-15-dATP. The fluorescein and tetramethylrhodamine tagged dATPs are suitable for the two-peak height encoded sequencing method. There is a mobility difference caused by the use of two different labels. This may be compensated for by suitable processing of the raw data.

The fluorescein-tetramethylrhodamine labelling pair suffers from high background in the tetramethylrhodamine channel. The spectral filter for the tetramethylrhodamine transmits water Raman scatter caused by the argon ion laser used to excite the fluorescein. Raman scatter degrades the signal-to-noise ratio for the

tetramethylrhodamine label; this label already gives less signal than the fluorescein.

Other label pairs, for example tetramethylrhodamine and Texas Red, or fluorescein and 5-carboxy-4',5'-dichloro-2',7'-dimethylfluorescein (JOE) would be better suited to this method. These dye pairs could be excited with one laser, the green helium neon or the blue argon ion, respectively, thus simplifying the instrument design. More importantly, the detection systems could be designed to minimise background and improve the signal-to-noise ratio for the labels, thus improving detection of small amounts of DNA.

References

1. Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H., Hood, L.M., *Nature*, **321** (1986), 674.
2. Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A., Bauermeister, K., *Science*, **238** (1987), 336.
3. Lee, L.G., Connell, C.R., Woo, S.L., Cheng, R.D., McArdle, B.F., Fuller, C.W., Halloran, N.D., Wilson, R.K., *Nucl. Acids Res.*, **20** (1992), 2471.
4. Voss, H., Schwager, C., Wirkner, U., Zimmermann, J., Erfle, H., Hewitt, N.A., Rupp, T., Stegemann, J., Ansorge, W., *Meth. Molec. Cell. Biol.*, **3** (1992), 30.
5. Voss, H., Wiemann, S., Wirkner, U., Schwager, C., Zimmermann, J., Stegemann, J., Erfle, H., Hewitt, N.A., Rupp, T., Ansorge, W., *Meth. Molec. Cell. Biol.*, **3** (1992), 153.
6. Hou, W., Smith, L.M., *Nucl. Acids Res.*, **21** (1993), 3331.
7. Chen, D.Y., Harke, H.R., Dovichi, N.J., *Nucl. Acids Res.*, **20** (1992), 4873.
8. Lu, H., Arriaga, A., Chen, D.Y., Dovichi, N.J., *J. Chromatogr.*, in press.
9. Tabor, S., Richardson, C.C., *Proc. Natl. Acad. Sci. USA*, **86** (1989), 4076.
10. Swerdlow, H.P., Zhang, J.Z., Chen, D.Y., Harke, H.R., Grey, R., Wu, S., Fuller, C., Dovichi, N.J., *Anal. Chem.*, **63** (1991), 2835.
11. H. Voss, K. Mühlegger, K. Effgen, S. Wiemann, U. Wirkner, J. Zimmermann, C. Schwager, W. Ansorge, *Biochemica*, **2** (1993), 8.

CHAPTER 6

Conclusions and Future Work

6.1 Conclusions

Capillary gel electrophoresis (CGE) is a powerful technique for sequencing DNA. Sequencing has been carried out with crosslinked polyacrylamide gels, and with linear polyacrylamide. The behaviour of short DNA fragments (< 250 bases) has been modelled in 5%C polyacrylamide gels ranging from 2.5 to 6%T. The observed mobilities, theoretical plate counts, and resolution of short fragments in gel-filled capillaries are similar to those in slab gels. While thermal gradients limit the performance of slab gels, the major source of band broadening in gel-filled capillaries is longitudinal diffusion. Further studies examining the effect of crosslinker concentration (%C) (1), electric field strength(2), and temperature (3) on the separation and behaviour of the DNA fragments have also been carried out independently of this work.

The two-colour peak-height method for DNA sequencing is a viable alternative to the existing labelling schemes and a number of different primer pairs may be used. The method works well for a variety of samples. In most cases the length of the sequence generated is limited by the resolving power of the gel rather than the signal-to-noise ratio of the sample. The reliability of the sequence late in the run is a function of both resolution and signal-to-noise ratio. In some cases, changing the coding scheme to suit the nature of the sample, *e.g.* A-T rich samples, increases the reliability of the data. The two-colour samples may also be separated on the automated ABI sequencer, although modifications to the software may be necessary. Unfortunately, the peak-height method is not currently compatible with cycle sequencing; the

thermostable enzymes required for cycle sequencing do not provide uniform incorporation of the ddNTPs.

A limited study of the performance of CGE suggests that the quality of the sequence generated is similar to that generated by commercially available, automated sequencers. The length and quality of the sequence obtained is in both cases highly dependent on the quality of the template DNA. Impure and degraded templates give ambiguous results. Compressions are major sources of error, but may be minimised by using nucleotide analogs (4), adding formamide to the gel (5), or by increasing the temperature at which electrophoresis is carried out (3). The main advantage of CGE over slab gel electrophoresis is the speed of separation.

Fluorescent deoxynucleotides (dNTPs) are an inexpensive alternative to fluorescent primers, particularly for primer walking applications. A labelling step is easily incorporated into a variety of sequencing protocols. Internal labelling in this fashion is compatible with peak-height encoding. Unfortunately, only fluorescein labelled nucleotides are currently available commercially.

6.2 Future Work

The main limitation of CGE to date is the sample throughput. The majority of systems in our laboratory use only one capillary *i.e.* only one sample at a time. Expansion of sample throughput is being addressed, however most multiple capillary systems are still limited to less than 20. Several samples may be analysed at once on a slab gel, whether one or four lanes are required per sample. Larger multiple capillary instruments are necessary for CGE to become a viable alternative to automated slab gel electrophoresis in routine sequencing applications.

All of the sequences analysed in this thesis were read by eye from the smoothed electropherograms. The development of appropriate computer programs to process the data and read the sequences is a logical extension of this work.

The internal labelling with dNTPs holds a great deal of promise. More labels are required before this method can be widely used with either two or four-colour labelling schemes. Of particular interest is the demonstration of internal labelling with cycle sequencing and with hexamer priming (6), where two or three contiguous hexamers are used in place of an 18mer to prime chain extension. Both methods should be within the capabilities of this laboratory.

Finally, the sequencing reactions, electrophoresis, and data acquisition are only a few of the steps necessary to sequence genomic DNA. Even though the automated slab gel sequencers are about an order of magnitude slower than CGE, they do not represent the bottleneck in terms of sequence generation. Analytical chemists may have a great deal to offer in the streamlining and automation of the steps leading up to the sequencing reactions.

References

1. D. Figeys, N.J. Dovichi, *J. Chromatogr.*, **645** (1993), 311.
2. N. Best, N.J. Dovichi, *J. Chromatogr.*,
3. H. Lu, E. Arriaga, D.Y. Chen, N.J. Dovichi, *J. Chromatogr.*, in press.
4. P.J. Barr, R.M. Thayer, P Laybourn, R.C. Najarian, F Seela, D.R. Tolan, *BioTechniques*, **4** (1986), 428.
5. M.J. Rocheleau, R. Grey, D.Y. Chen, H.R. Harke, N.J. Dovichi, *Electrophoresis*, **13** (1992), 484-486.
6. J. Kieleczawa, J.J. Dunn, F.W. Studier, *Science*, **258** (1992), 1787.

Appendix A

Consensus Sequences

1. M13mp18 - the universal primer site (-21) is underscored.

	10	20	30	40	50	60	
1	AACATCCAAT	AAATCATACA	GGCAAGGCAA	AGAATTAGCA	AAATTAAGCA	ATAAAGCCTC	60
61	AGAGCATAAA	GCTAAATCGG	TTGTACCAAA	AACATTATGA	CCCTGTAATA	CTTTTGCGGG	120
121	AGAAGCCTTT	ATTTCAACGC	AAGGATAAAA	ATTTTATAGAA	CCCTCATATA	TTTAAAAGC	180
181	AATGCCTGAG	TAATGTGTAG	GTAAGATTTC	AAAAGGGTGA	GAAAGGCCGG	AGACAGTCAA	240
241	ATCACCATCA	ATATGATATT	CAACCGTTCT	AGCTGATAAA	TTAATGCCGG	AGAGGGTAGC	300
301	TATTTTGTAG	AGATCTACAA	AGGCTATCAG	GTCATTGCCT	GAGAGTCTGG	AGCAAACAAG	360
361	AGAATCGATG	AACGGTAATC	GTAAAACCTAG	CATGTCAATC	ATATGTACCC	CGGTTGATAA	420
421	TCAGAAAAGC	CCCAAAACA	GGAAAGATTGT	ATAAGCAAAT	ATTTAAATFG	TAAACGTTAA	480
481	TATTTTGTTA	AAATTCGCGT	TAAATTTTGT	TTAAATCAGC	TCATTTTPTA	ACCAATAGGA	540
541	ACGCCATCAA	AAATAATTCG	CGTCTGGCCT	TCCTGTAGCC	AGCTTTCATC	AACATTAAAT	600
601	GTGAGCGAGT	AACAACCCGT	CGGATTCTCC	GTGGGAACAA	ACGGCGGATT	GACCGTAATG	660
661	GGATAGGTTA	CGTTGGTGTA	GATGGGCGCA	TCGTAACCGT	GCATCTGCCA	GTTTGAGGGG	720
721	ACGACGACCG	TATCGGCCTC	AGGAAGATCG	CACTCCAGCC	AGCTTTCGGG	CACCGTTCT	780
781	GGTGCCGAA	ACCAGGCAAA	GCGCCATTTCG	CCATTCAGGC	TCCGCAACTG	TTGGGAAGGG	840
841	CGATCGGTGC	GGCCCTCTTC	GCTATTACGC	CAGCTGGCGA	AAGGGGGATG	<u>AGCTGCAAGG</u>	900
901	CGATTAAAGT	GGTAACGCC	AGGGTTTTC	CAGTCACGAC	<u>GTGTAAAAC</u>	<u>GACGGCAAT</u>	960
961	GCCAAGCTTG	CATGCCTGCA	GGTCGACTCT	AGAGGATCCC	CGGGTACCGA	GCTCGAATTC	1020
1021	GTAATCATGG	TCATAGCTGT	TTCTGTGTG	AAATGTTTAT	CCGCTCACA	TTCCACACA	1080
1081	CATACGAGCC	GGAAACATAA	AGTGTAAGC	CTGGGGTGCC	TAATGAGTGA	GCTAACTCAC	1140
1141	ATTAATTGCG	TTGCGCTCAC	TGCCCGCTTT	CCAGTCGGGA	AACCTGTCGT	GCCAGCTGCA	1200
1201	TTAATGAATC	GGCCAACCGG	CGGGGAGAGG	CGGTTTGCCT	ATTGGGCGCC	AGGGTGGTTT	1260
1261	TTCTTTTCAC	CAGCGAGACG	GGCAACAGCT	GATTGCCCTT	CACCGCCTGG	CCCTGAGAGA	1320
1321	GTTGCAGCAA	GCGGTCCACG	CTGGTTTGCC	CCAGCAGGCG	AAAATCCTGT	TTGATGGTGG	1380
1381	TTCCGAAATC	GGCAAAATCC	CTTATAAATC	AAAAGAATAG	CCCAGATAG	GGTTGAGTGT	1440
1441	TGTTCCAGTT	TGGAACAAGA	GTCCACTATT	AAAGAPCGTG	GACTCCAACG	TCAAAGGGCG	1500
1501	AAAAACCGTC	TATCAGGGCG	ATGGCCCACT	ACGTGAACCA	TCACCCAAAT	CAAGTTTTTT	1560
1561	GGGGTCGAGG	TGCCGTAAG	CACTAAATCG	GAACCCATAA	GGGAGCCCC	GATTTAGAGC	1620
1621	TTGACGGGGA	AAGCCGGCGA	ACGTGGCGAG	AAAGGAAGGG	AAGAAAGCGA	AAGGAGCGGG	1680
1681	CGCTAGGGCG	CTGGCAAGTG	TAGCGGTAC	GCTGCCGTA	ACCACCACAC	CCGCCGCGCT	1740
1741	TAATGCGCCG	CTACAGGGCG	CGTACTATGG	TTGCTTTGAC	GAGCACGTAT	AACGTGCTTT	1800
1801	CCTCGTTGGA	ATCAGAGCGG	GAGCTAAACA	GGAGGCCGAT	TAAAGGGATT	TTAGACAGGA	1860
1861	ACGGTACGCC	AGAATCTTGA	GAAGTGTTTT	TATAATCAGT	GAGGCCACCG	AGTAAAAGAG	1920
1921	TCTGTCCATC	ACGCAAATTA	ACCGTTGTAG	CAATACTTCT	TGATTAGTA	ATAACATCAC	1980
1981	TTGCCCTGAGT	AGAAGAATCC	ATCTATCGG	CCTTGCTGGT	AAATCCAGA	ACAATATTAC	2040
2041	CGCCAGCCAT	TGCAACAGGA	AAACGCTCA	TGGAATACC	TACATTTTGA	CGCTCAATCG	2100
2101	TCTGAAATGG	ATTATTTTACA	TTGGCAGATT	CACCAGTCAC	ACGACCAGTA	ATAAAGGGA	2160
2161	CATTCTGGCC	AACAGAGATA	GAACCCTTCT	GACCTGAAAG	CTAAGAATA	CGTGGCACAG	2220
2221	ACAATATTTT	TGAATGGCTA	TTAGTCTTTA	ATGCCGGAAC	TGATAGCCCT	AAAACATCCG	2280
2281	CATTAATAAT	ACCGAACGAA	CCACCAGCAG	APGATAAAAC	AGAGGTGAGG	CGGTCAGTAT	2340
2341	TAACACCGCC	TGCAACAGTG	CCACGCTGAG	AGCCAGCAGC	AAATGAAAAA	TCTAAAGCAT	2400
2401	CACCTTGCTG	AACCTCAAAT	ATCAAACCCT	CAATCAATAT	CTGGTCAGTT	GGCAAATCAA	2460
2461	CAGTAGAAAG	GAATTGAGGA	AGGTTATCTA	AAATACTTTT	AGGTGCACTA	ACAATAATA	2520
2521	GATTAGAGCC	GTCAATAGAT	AATACATTTG	AGGATTTAGA	AGTATTAGAC	TTTACAAACA	2580
2581	ATTCGACAAC	TGTTATTTAA	TCCTTTGCC	GAACGTTATT	AAATTTAAA	GTTTGAGTAA	2640
2641	CATTATCATT	TTGCCGAACA	AAGAAACCAC	CAGAAGGAGC	GGAATTATCA	TCATATTCCT	2700
2701	GATTATCAGA	TGATGGCAAT	TCATCAATAT	AATCCTGATT	GTTTGATTAA	TACTTCTGAA	2760
2761	TTATGGAAGG	AATTGAACCA	ACCATATCAA	AATTATTAGC	ACGTAAAACA	GAAATAAAGA	2820
2821	AATTGCGTAG	ATTTTCAGGT	TTAACGTCAG	ATGAATATAC	AGTAACAGTA	CCTTTTACAT	2880
2881	CGGGAGAAAC	AATAACGGAT	TCCGCTGATT	GCTTTGAATA	CCAAGTTACA	AAATCCCGCA	2940
2941	GAGGCGAATT	ATTCATTTCA	ATTACCTGAG	CAAAGAAGA	TGATGAAACA	AACATCAAGA	3000

3001	AAACAAAATT	AATTACATTT	AACAATTTCA	TTTGAATTAC	CTTTTTTAAT	GGAACAGTA	3060
3061	CATAAATCAA	TATATGTGAG	TGAATAACCA	TGCTTCTGTA	AATCGTGGCT	ATTAATTAAT	3120
3121	TTTCCCTTAG	AATCCTTGAA	AACATAGCGA	TAGCTTAGAT	TAAGACGCTG	AGAAGAGTCA	3180
3181	ATAGTGAATT	TATCAAAATC	ATAGGTCTGA	GAGACTACCT	TTTTAACCTC	CGGCTTAGGT	3240
3241	TGGGTATAT	AACTATATGT	AAATGCTGAT	GCAAATCCAA	TGCAAGACA	AAGAACCGCA	3300
3301	GAAAACCTTT	TCAAATATAT	TTTAGTTAAT	TTTCATCTCT	GACCTAAATT	TAATGGTTTG	3360
3361	AAATACCGAC	CGTGTGATAA	ATAAGGCGTT	AAATAAGAAT	AAACACCGGA	ATCATAATTA	3420
3421	CTAGAAAAAG	CCTGTTTAGT	ATCATATGCG	TTATACAAAT	TCTTACCAGT	ATAAAGCCAA	3480
3481	CAATCAACAG	TAGGCTTAA	TTGAGAATCG	CCATATTTAA	CAACGCCAAC	ATGTAATTTA	3540
3541	GGCAGAGGCA	TTTTCGAGCC	AGTAATAAGA	GAATATAAAG	TACCGACAAA	AGGTAAAGTA	3600
3601	ATTCTGTCCA	GACGACGACA	ATAACAACA	TGTTGAGCTA	ATGCAGAACG	CGCCTGTTTA	3660
3661	TCAACAATAG	ATAATCTCTG	AACAAGABAA	ATAATATCCC	ATCTAATTTT	ACGAGLATT	3720
3721	AGAAACCAAT	CAATAATCGG	CTGTCTTTCC	TTATCATTCC	AAGAACGGGT	ATTAACCAA	3780
3781	GTACCCACT	CATCGAGAAC	AAGCAAGCCG	TTTTTATTTT	CATCGTAGGA	ATCATTACCG	3840
3841	CGCCCAATAG	CAAGCAAATC	AGATATAGAA	GGCTTATCCG	GTATTCTAAG	AACGCGAGGC	3900
3901	GTTTTAGCGA	ACCTCCCGAC	TTGCGGGAGG	TTTTGAAGCC	TTAAATCAAG	ATTAGTTGCT	3960
3961	ATTTTGCACC	CAGCTACAAT	TTTATCCTGA	ATCTTACCAA	CGTAACGAG	CGTCTTTCCA	4020
4021	GAGCCTAATT	TGCCAGTTAC	AAAATAAACA	GCCATATTAT	TTATCCCAAT	CCAAATAAGA	4080
4081	AACGATTTTT	TGTTTAACTG	CAAAAATGAA	AATAGCAGCC	TTTACAGAGA	GAATAACATA	4140
4141	AAAACAGGGA	AGCGCATTAG	ACGGGAGAAT	TAAGTGAACA	CCCTGAACAA	AGTCAGAGGG	4200
4201	TAATTGAGCG	CTAATATCAG	AGAGATAACC	CACAAGAATT	GAGTTAAGCC	CAATAATAAG	4260
4261	AGCAAGAAAC	AATGAAATAG	CAATAGCTAT	CTTACCGAAG	CCCTTTTTAA	GAAAAGTAAAG	4320
4321	CAGATAGCCG	AACAAAGTTA	CCAGAAGGAA	ACCGAGGAAA	CGCAATAATA	ACGGAATACC	4380
4381	CAAAAGAACT	GGCATGATTA	AGACTCCTTA	TTACGCAGTA	TGTTAGCAAA	CGTAGAAAAT	4440
4441	ACATACATAA	AGGTGGCAAC	ATATAAAAGA	AACGCAAAGA	CACCACGGAA	TAAGTTTATT	4500
4501	TTGTCACAAT	CAATAGAAAA	TTTATATGGT	TTACCAGGCG	TAAAGACAAA	AGGGCGACAT	4560
4561	TCAACCGATT	GAGGGAGGGA	AGGTAAATAT	TGACGGAAAT	TATTCATTA	AGGTGAATTA	4620
4621	TCACCGTCAC	CGACTTGAGC	CATTTGGGAA	TTAGAGCCAG	CAAAATCACC	AGTAGCACA	4680
4681	TTACCATTAG	CAAGGCCGGA	AACGTACCA	ATGAAACCAT	CGATAGCAGC	ACCGTAATCA	4740
4741	GTAGCGACAG	AATCAAGTTT	GCCPTTAGCG	TCAGACTGTA	GCGCGTTTTC	ATCGGCATTT	4800
4801	TGGTTCATAG	CCCCCTTATT	AGCGTTTGCC	ATCTTTTCAT	AATCAAAATC	ACCGGAACCA	4860
4861	GAGCCACCAC	CGGAACCGCC	TCCCTCAGAG	CCGCCACCCT	CAGAACCGCC	ACCCTCAGAG	4920
4921	CCACCACCCT	CAGAGCCGCC	ACCAGAACCA	CCACCAGAGC	CGCCGCCAGC	ATTGACAGGA	4980
4981	GGTTGAGGCA	GGTCAGACGA	TTGCCTTGA	TATTCACAAA	CGAATGGATC	TTCATTAAG	5040
5041	CCAGAATGGA	AAGCGCAGTC	TCTGAATTTA	CCGTTCCAGT	AAGCGTCATA	CATGGCTTTT	5100
5101	GATGATACAG	GAGTGTACTG	GTAATAAGTT	TTAACGGGGT	CAGTGCCTTG	AGTAACAGTG	5160
5161	CCCGTATAAA	CAGTTAATGC	CCCTGCCTA	TTTCGGAAAC	TATTATCTGT	AAACATGAAA	5220
5221	GTATTAAGAG	GCTGAGACTC	CTCAAGAGAA	GGATTAGGAT	TAGCGGGGTT	TTGCTCAGTA	5280
5281	CCAGGCCGAA	AGGTGCGGTC	GAGAGGGTTG	ATATAAGTAT	AGCCCGGAAT	AGGTGTATCA	5340
5341	CCGTAATCAG	CAGGTTTAGT	ACCGCCACCC	TCAGAACCAG	CACCCTCAGA	ACCGCCACCC	5400
5401	TCAGAGCCAC	CACCCTCATT	TTCAGGGATA	GCAAGCCCAA	TAGGAACCCA	TGTACCCTAA	5460
5461	CACGTAGTTT	CGTCACCAGT	ACAAACTACA	ACGCCCTGAG	CATTCCACAG	ACAACCCTCA	5520
5521	TAGTTAGCGT	AACGATCTAA	AGTTTGTGCG	TCTTTCCAGA	CGTTAGTAAA	TGAATTTTCT	5580
5581	GTATGGGGTT	TTGCTAAACA	ACTTTCAACA	GTTTCAGCGG	AGTGAGAATA	GAAAGGAACA	5640
5641	ACTAAAGGAA	TTGCGAATAA	TAATTTTTTC	ACGTTGAAAA	TCTCCAAAAA	AAAAGGCTCC	5700
5701	AAAAGGAGCC	TTTAATTGTA	TCGGTTTATC	AGCTTGCTTT	CGAGGTGA	TTCTTAAACA	5760
5761	GCTTGATACC	GATAGTTGCG	CCGACAATGA	CAACAACCAT	CGCCACGCA	TAACCGATA	5820
5821	ATTCGGTTCG	TGAGGCTTGC	AGGGAGTTAA	AGGCCGCTTT	TGCGGGATCG	TCACCCTCAG	5880
5881	CAGCGAAGA	CAGCATCGGA	ACGAGGGTAA	CAACGGCTAC	AGAGGCTTTG	AGGACTAAAG	5940
5941	ACTTTTTTCAT	GAGGAAGTTT	CCATTAACG	GGTAAAATAC	GTAATGCCAC	TACGAAGGCA	6000
6001	CCAACCTAAA	ACGAAAGAGG	CGAAAGAATA	CACTAAAACA	CTCATCTTTG	ACCCCCAGCG	6060
6061	ATTATACCAA	GCGCGAACA	AAGTACAACG	GAGATTTGTA	TCATCGCCTG	ATAAATTGTG	6120
6121	TCGAAATCCG	CGACCTGCTC	CATGTTACTT	AGCCCGAACG	AGGCCGAGAC	GGTCAATCAT	6180
6181	AAGGGAACCG	AACTGACCAA	CTTTGAAAGA	GGACAGATGA	ACGGTGTACA	GACCAGCGCG	6240
6241	ATAGGCTGGC	TGACCTTCAT	CAAGAGTAAT	CTTGACAAGA	ACCGGATATT	CATTACCCAA	6300
6301	ATCAACGTAA	CAAAGCTGCT	CATTCAGTGA	ATAAGGCTTG	CCCTGACGAG	AAACACCAGA	6360
6361	ACGAGTAGTA	AATTGGGCTT	GAGATGGTTT	AATTTCAACT	TTAATCATTG	TGAATTAACCT	6420
6421	TATGCGATTT	TAAGAACCTG	CTCATTATAC	CAGTCAGGAC	GTTGGGAAGA	AAAATCTACG	6480
6481	TTAATAAAAC	GAATAACCG	AACAACATTA	TTACACGTAG	AAAGATTCAT	CAGTTGAGAT	6540
6541	TTAGGAATAC	CACATTCAAC	TAATGCAGAT	ACATAACGCC	AAAAGGAATT	ACGAGGCATA	6600
6601	GTAAGAGCAA	CACATTCATA	ACCCTCGTTT	ACCAGACGAC	GATAAAAACC	AAAATAGCGA	6660

6661	GAGGCTTTTG	CAAAGAAGT	TTTGCCAGAG	GGGTAATAG	TAAAATGTTT	AGACTGGATA	6720
6721	GCGTCCAATA	CTGCGGAATC	GTCATAAATA	TTCATTGAAT	CCCCCTCAA	TGCTTTAAAC	6780
6781	AGTTCAGAAA	ATGAGAATGA	CCATAAATCA	AAAATCAGGT	CTTTACCCTG	ACTATTATAG	6840
6841	TCAGAAGCAA	AGCGGATTGC	ATCAAAAAGA	TTAAGAGGAA	GCCCGAAAGA	CTTCAAATAT	6900
6901	CGCGTTTAA	TTCGAGCTTC	AAAGCGAACC	AGACCGGAAG	CAAACCCAA	CAGGTCAGGA	6960
6961	TTAGAGAGTA	CC...NTG	CTCCTTTTGA	TAAGAGGTCA	TTTTTGCGGA	TGGCTTAGAG	7020
7021	CTTAATGCT	GAATCTGGTG	CTGTAGCTCA	ACATGTTTTA	AATATGCAAC	TAAAGTACGG	7080
7081	TGTCTGGAAG	TTTCATTCCA	TGTAACAGTT	GATTCCCAAT	TCTGCGAACG	AGTAGATTTA	7140
7141	GTTTGACCAT	TAGATACATT	TCGCAAATGG	TCAATAACCT	GTTTAGCTAT	ATTTTCATTT	7200
7201	GGGCGCGGAG	CTGAAAAGGT	GGCATCAATT	CTACTAATAG	TAGTAGCATT		7250
	10	20	30	40	50	60	

2. Mouse mammary tumour virus (MMTV)

	10	20	30	40	50	60	
1	GTTTACATAA	GCATTTACAT	AAGACTTGA	TAAGTTCCAA	AAGAACATAG	GAGAATAGAA	60
61	CATTCAGAGC	TTAGATCAAA	ACATTTGATA	CCAAACCAAG	TCAGGAAACT	ACTTGTCTCA	120
121	CATCCTTGCA	CCTGTTCTTC	AATTGAGGTT	GAGCGTCTCT	TTCTATTTTC	TATCCCATT	180
181	TCTAACTTCT	GAATTTGAGT	AAAATAGTA	CTAAAAGATA	ATGATTCAT	TCTTAACATA	240
241	GTAACATAA	ATCTACCTAT	TGGATTGGTC	TTATTGGTAA	AAATATAATT	TTAGCAAGC	300
301	ATTCTTATTT	CTATTTCTGA	AGGACAAAGT	CGGTGTGGCT	TGTAAGAGGA	AGTTGGCTGT	360
361	GGTCCTTGCC	CCAGGAGGAA	GTTGAGTTC	TCCGAATCGT	TTAGATTGTA	ATCTGCACA	420
421	GAAGAGTAAT	TAAAAGAATC	AAGGGTGAGA	GCCCTGCGAG	CACGAACCGC	AACTTTCCCC	480
481	AATAGCCCCA	GGCAAAGCAG	AGCTATGCCA	AGTTTGCAGC	AGAGAATGAA	TATGTCTTTG	540
541	TCTGATGGGC	TCATCCGTTT	GTGCGCAGAC	AGGTGCTCCT	TGGTGGGAAA	CAACCCCTTG	600
601	GCTGCTTCTC	CCCTAGGTGT	AGGACACTCT	CGGGAGTTCA	ACCATTTCTG	CTGCAGGCGC	660
661	GGCATTTCCC	CCTTTTTTCT	TTTTTAAAS	AAGCACGTTA	AGATCTGACT	GCACTTGGTC	720
721	AAGGCTCTTC	GCAAGGCACT	GGAAAACAAT	GGGGAAAATC	ATAAGTACTA	TGACCAAAG	780
781	CAGGGCTCCA	ACTCCTATAA	AAATGAAATA	TTGTGTCCAA	TCCAATGGAT	TTAAAGCCTT	840
841	TACTCCATTG	GCAAAGGACT	GAGCCAAGCT	ACTGAGGTCC	ACTGCGTCAA	TATGTTGTTT	900
901	GCTCATATCA	CTAATCAGGT	TGGTAACTC	CTGTATGTTA	TATGAAATCT	TATTGTCATT	960
961	CCAAATGCCC	AATAAATGTG	CTCTGGTTCT	TTCCCAGCTC	TCAGAAGCAT	TATATGGCAA	1020
1021	AGGTGTGACA	CAGATAAAAT	CATGATTTGC	ATGACACCTA	GTGGACATTC	TGGTCTTTAA	1080
1081	GTTTGCCACA	TCTTGACCCA					1100
	10	20	30	40	50	60	

3. Mouse cytokine - 123 delta

	10	20	30	40	50	60	
1	gccttgc	atg cctgcaggtc	gactCTAGAC	CACCATGGCG	CTCTGGGTGA	CTGCAGTCCT	60
61	GGCTCTTGCT	TGCCCTGGTG	GTCTCGCCGC	CCCAGGGCCG	GTGCCAAGAT	CTGTGTCTCT	120
121	CCCTCTGACC	CTTAAGGAGC	TTATTGAGGA	GCTGAGCAAC	ATCACACAAG	ACCAGACTCC	180
181	CCTGTGCAAC	GCGAGCATGG	TATGGAGTGT	GGACCTGGCC	GCTGGCGGGT	TCTGTGTAGC	240
241	CCTGGATTCC	CTGACCAACA	TCTGCAATTG	CAATGCCATC	TACAGGACCC	AGAGGATATT	300
301	GCATGGCCTC	TGTAACCGCA	AGGCCCCAC	TACGGTCTCC	AGCCTCCCCG	ATACCAAAT	360
361	CGAAGTAGCC	CACTTTATAA	CAAACTGCT	CAGCTACACA	AAGCAACTGT	TTCGCCACGG	420
421	CCCCTTCTAA	GCGGCCGCGG	AGGCCGAATT	CCGTCGAGGG	ATCC		464
	10	20	30	40	50	60	

4. Malaria templates - clones 1 to 5

Clone 1

1	CTTCATTTTT	TATTGTA	TCATAAGTAT	CACCATCTTG	GTCTTCATAA
51	TATTCGTCAA	CTCTTATCAT	ATCTCCATCC	CGTGCTTCTA	CAG_AATCTT
101	CATCCTCATC	ATTA	TCTTCTGAAG	ATTCAGCACC	TACTTCTTCC
151	TGATGTTCCCT	CTTGTAATTG	TTCTTCTGTT	TCATCCCCAC	TTTCTTCATA
201	AACATAATCA	TCATATICTT	CCTCCTCCTC	TTCTTCTTCT	TCTTCCTCTT
251	CTTCCTCTTC	CTCTTCTTCT	TCCTCCTCCT	CTTCTTCCTC	ATCATCTATT
301	ACAAAATTAT	CCCTCGGCGA	CATGTCTTCC	TCATCCAAAT	TTCTGTTTCG
401	TCAACGAACT	GTTTATAAAA	AAATTAGACT	TATGTTTTTT	TTGTACTACA
451	GTTTTTTTTTC	TTTTGGGTGA			

Clone 2

1	TCCTCATCAT	CTATTACAAA	ATTATCCCTG	GGCGACATGT	CTTCCTCATC
51	CAAATTTCTG	TTTCGACGTA	CTAGGTAAAA	GCTATCCGAG	GGGGTTAAAT
101	ACATATATAT	ATATTTCAAC	GAAGTGTTTA	TAAAAAAATT	AGACTTATGT
151	TTTTTTTGTA	CTACAGTTTT	TTTTCTTTTG	GGTGAAGAAA	ATATTTTCAT
201	TTTTTTTGTA	CTACAGTTTT	TTTTCTTTTG	GGTGAAGAAA	ATATTTTCAT
251	CATCAAACCA	TTCTATAAGT	GTATCATCAC	TTTCATAATC	TTTATTTTCG
301	CATCAAACCA	TTCTATAAGT	GTATCATCAC	TTTCATAATC	TTTATTTTCG
351	TTCTTTATCT	ATATTTCTTA	TTAATATCAT	ATTTTTTTCTT	GTACATTGGA
401	ATCGATATCA	CTACTAACAA	CATGATCATT	TGTTACATTC	TTTTTTATAA
451	TAATTGTGTC	CAGTGTTTTC	CTTTCCTTCT	TCCCTACGAT	

Clone 3

1	TTCTAATTAT	TTTTATAACA	TTTTCTCCAT	CTTCGACAGC	TTCACCATAT
51	TCACCATATT	CACCATCTCG	TGCAGTATGT	TTTTTAGATA	CTCTACTAAT
101	TGTATCATCA	TCTAATAAAA	CTTTGGAACC	ATCCACTTTA	TATACATCAT
151	CTCCTTCACC	CCCTTCTATA	AATTTAAACA	AATCTGTTTT	ATCAACCTCG
201	CTCTTTTCAT	AGGTAAAAC	TTCTCCACCT	TCATCTTTTTT	CTTCCTCATC
251	ACCAAATGGA	TATATTCAC	CTTGCCCTTC	TTTTTCATCT	ACATATTCAC
301	CTTCTTCTTC	ACCTACCTCT	TCACCTTCTT	CTTCACCTAC	CT

Clone 4

1	TAAAAAAAAAT	TGGGATGATG	TGGGATGATG	TACATTTATT	TTATCCTCCT
51	CATAATGTAT	TACATAATGT	TGTACTTAAT	AATCATATAG	TCAACTTATC
101	ATCTGCATTA	GAAGGAGTCT	TATTTATGAA	ATCAAAAAGTT	ACTGGAGATG
151	AAACAGCTAC	AAAAAAAAAAC	ACTACACTAC	CAACTGATGG	TGTATCAAGT
201	ATTTTAATTC	CACCATATGT	AAAGGAAGAT	ATAACATTTT	ATCTTTTTTTG
251	TGGGAAATCT	ACAACAAAAA	AACCAAACAA	AAAGAACACA	AACAAATGAT
301	GCTATTAGTA	ATAATAATAA	TAATTCATAT	TCTATATTTA	CACATAATAA
351	AAATACAGAG	AATAATCTAA	TATGTGATAT	ATCTTTAATT	CCAAAAACTG
401	TTATAGGAAT	TAAATGTCCT	AATAAAAAAT	TAAATCCACA	AACATGTTTT
451	GATGAAGTGT	ATTATGTTAA	ACAAGAAGAT	GTACCTTCGA	AAACTATAAC

Clone 5

1	AAATATAAAAA	ATTTTCATTA	AAACCATCAT	TAGTTTTTGA	TGATAACAAT
51	CTATTAGTTT	AGCTTTGAAA	GGGGTTTATG	GAAATCGAAT	TTTTACTTTT
101	GATAAAAATG	GAAAAAAAGG	AGAAGGAATT	AGTTTTTTTTA	TACCTCCAAT
151	AAAACAAGAT	ACAGATTTAA	AATTTATAAT	TAATGAAACA	ATAGATAATT
201	CAAATATTAA	ACAAAGAGGA	TTAATATATA	TTTTTGTTAG	GAAAAATGTA
251	TCAGAAAATT	CATTTAAATT	ATGTGATTC	ACAACAGGTC	GACTTCATTA
301	ATGGAATTAA	ATAGTCAAGT	AAAAGAAAA	AACTTGCACT	GTAAAATTA
351	AAAAAGGAGA	TATTTTGGGA	TTGAAATCTC	CTAAAGGTTT	TGCTATATTT
401	CCACAAGCAT	GTTTTAGTAA	TGTTTTATTA		

Appendix B

Problems in DNA Sequencing

Things That Go Bump in the Night

This appendix is intended as a guide to a few of the well known problems encountered in DNA sequencing. The illustrative data here are all capillary electropherograms. Electropherograms resemble chromatograms more than the usual autoradiograms, and the bulk of the literature that deals with "things that go bump in the night" is illustrated with autoradiograms. The three ~~examples~~ that follow have all been encountered in the course of the work described in ~~the thesis~~.

1. Compressions

Compressions occur when a DNA fragment folds back and base pairs with itself. This happens more commonly in regions with a high G-C content. The fragment then migrates as if its length were several bases shorter, and coelutes with shorter fragments, giving what appears to be a pile-up of peaks in the electropherogram. A typical compression is shown in Figure B.1. Compressions are a problem because they make the electropherogram difficult to read and introduce errors or ambiguities into the sequence.

There are two approaches to resolving compressions: The first approach is a chemical approach, where nucleotide analogs e.g. 7-deazadideoxyguanosine triphosphate or inosine triphosphate are used in place of dGTP(1-2). The base pairing interactions between these analogs and C in the DNA strand are weaker than the G-C base pairs, making the denaturing conditions of a typical slab gel adequate to resolve many compressions. This approach has not been widely implemented in our

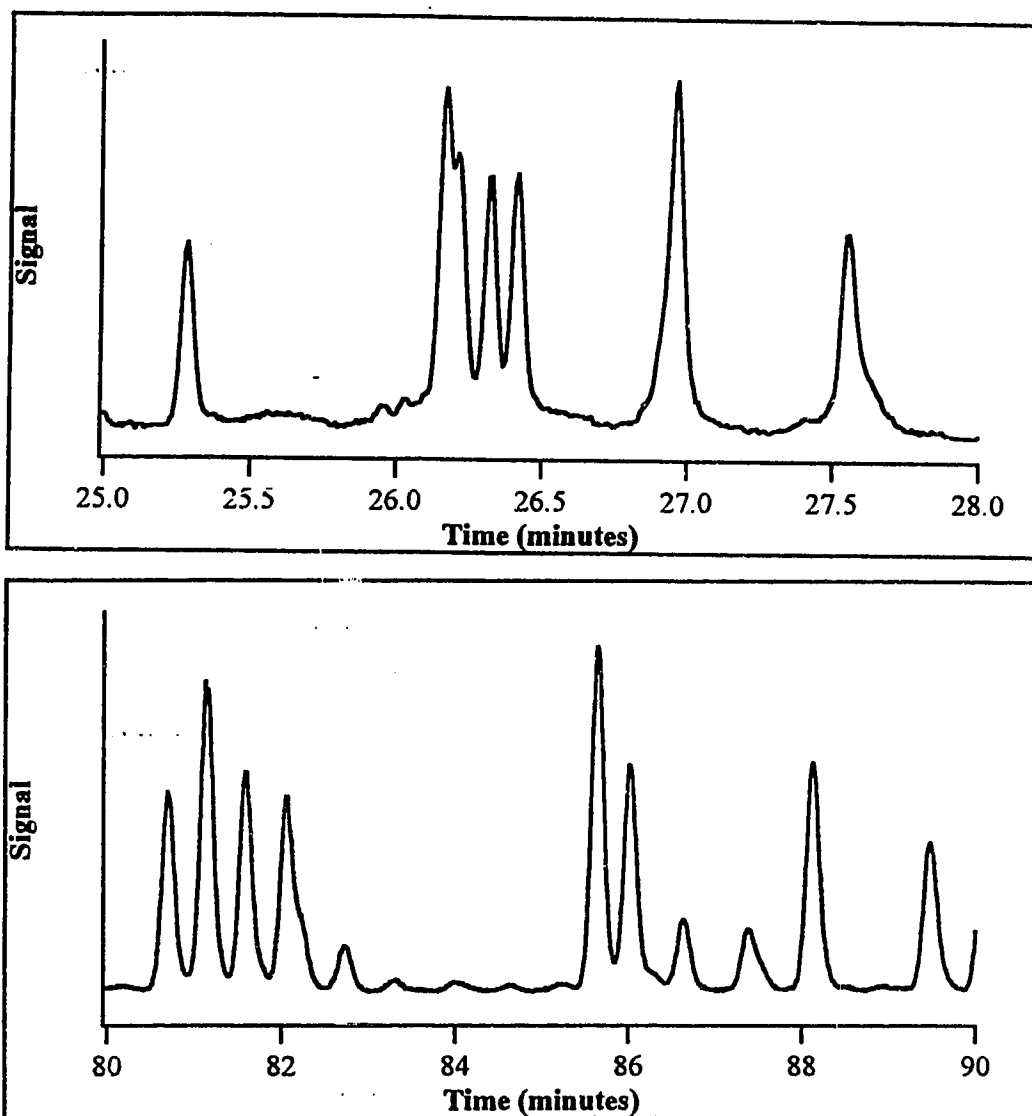


Figure B.1 The effect of formamide on compressions. The top panel shows compressions in the first two of the quartet of C's (at 26.2 min) and at the doublet (26.9 min). The sample was electrophoresed on a 4%T 5%C gel. The bottom panel shows a C terminated M13mp18 sample was electrophoresed on a 4%T 5%C gel with 10% (v/v) formamide added. The compressions are completely resolved (the small peaks are due to template degradation).

laboratory.

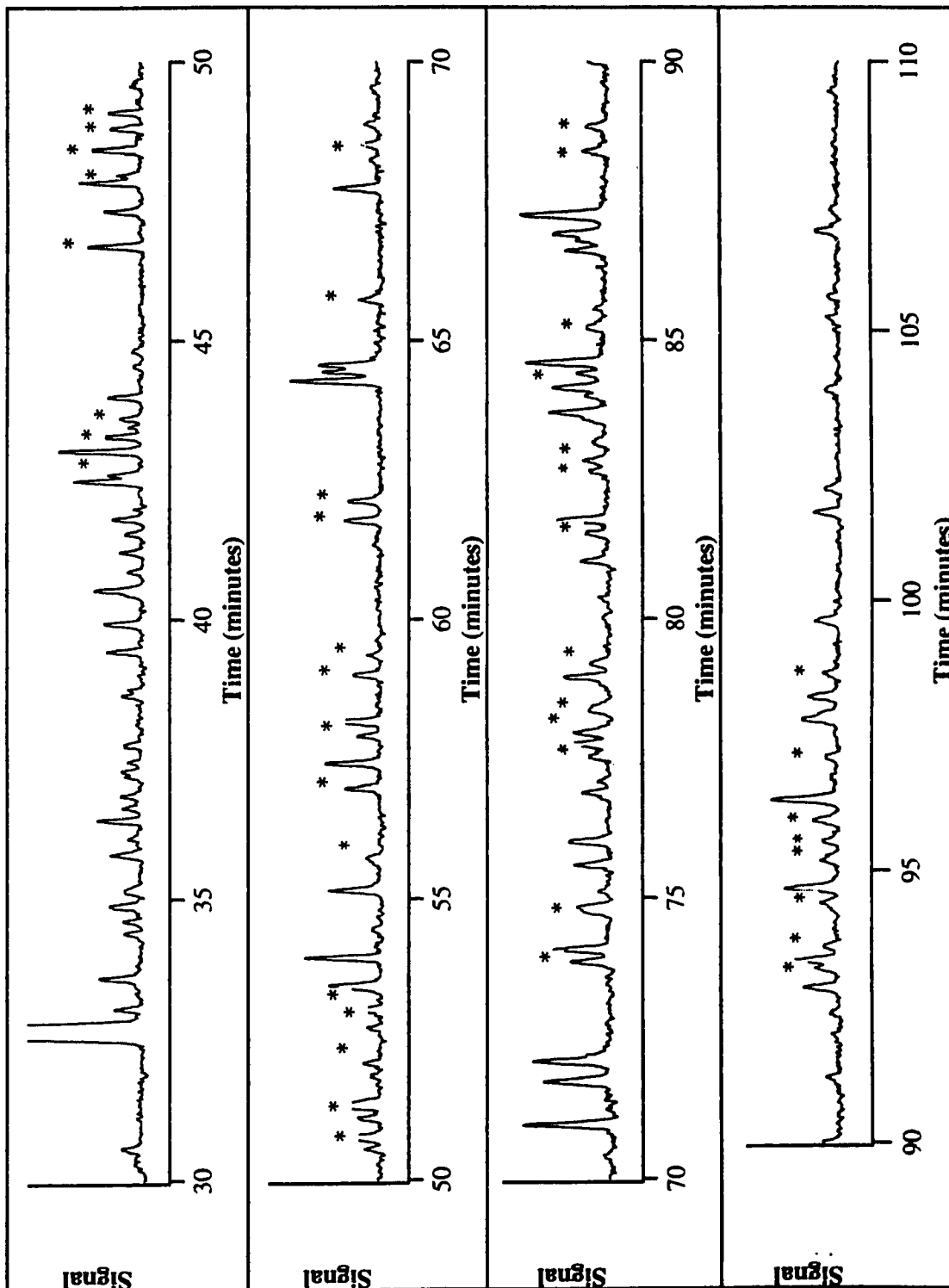
The second approach involves manipulating the conditions of the electrophoretic separation to prevent compressions from occurring. Two different sets of conditions help resolve compressions: 1. the addition of formamide to the gel, and 2. carrying out electrophoresis at an elevated temperature. Formamide is a chaotropic agent that disrupts the hydrogen bonds that form the base pairs. The effect of formamide on a compression is illustrated in Figure B.1. While formamide helps resolve compressions, it also slows the migration of the DNA fragments through the gel. Rocheleau *et al.* (3) determined that the optimum concentration of formamide in the gel is 10% (v/v). Above the optimum concentration only a minimal gain in resolution is obtained, while the sequencing rate decreases dramatically.

Thermostating the capillaries at temperatures above room temperature also reduces compressions as shown by Lu *et al.* (4) simply because there is sufficient thermal energy to break the hydrogen bonds. Of the two methods, thermostating the capillaries is generally preferred, since higher temperature also results in faster sequencing rates.

2. Non-specific Priming

The chain termination method of DNA sequencing relies on the specificity of the annealing step. If the primer is complementary to the sequence in more than one place along the vector or the insert, it can anneal and be extended, resulting in a mixture of fragments. Non-specific priming can also occur where several bases at the 3' end of the primer (the end that is extended by the polymerase) are complementary to a portion of the template. This is more likely to happen if the concentration of primer is high. When there is a large amount of primer relative to template the small fraction

Figure B.2 Non-specific priming. The electropherogram of an A terminated M13mp18 sample. The peaks marked with asterisks do not correspond to A's and are due to non-specific priming. These peaks disappeared when the amount of primer was decreased.



of non-specific priming events yields a detectable amount of product. Such a case is shown in figure B.2. The sample was A-terminated M13mp18 DNA, and 30 ng of primer was added to 1 µg of M13mp18 template. The peaks marked with asterisks do not correspond to the A terminated fragments expected for this sample. When the amount of primer was reduced to 20 ng the extra peaks disappeared from the electropherogram.

3. Degraded Template DNA

Over time, a solution of template DNA may degrade, particularly if it is subjected to frequent freeze-thaw cycles. When degradation occurs the DNA is cleaved in one or more positions, resulting in fragments of various lengths. When the chain extension reaction is carried out, extension is terminated if 1. a ddNTP is incorporated into the chain or 2. if the template molecule is cleaved. DNA fragments resulting from both of these events are detected when the samples are prepared with labelled primers. Figure B.3a shows a C terminated M13mp18 sample prepared with the FAM labelled primer. In addition to the peaks which correspond to the C terminated fragments there are many other peaks. In particular there are several high intensity peaks at about 94 min, 142 min, and 146 min that do not correspond to C-terminated fragments. These are due to the second type of termination, where the template molecules are cleaved. Figure B.3b shows an A terminated sample that also exhibits evidence of template degradation.

Figure B.3 Template degradation
a. A C terminated sample.

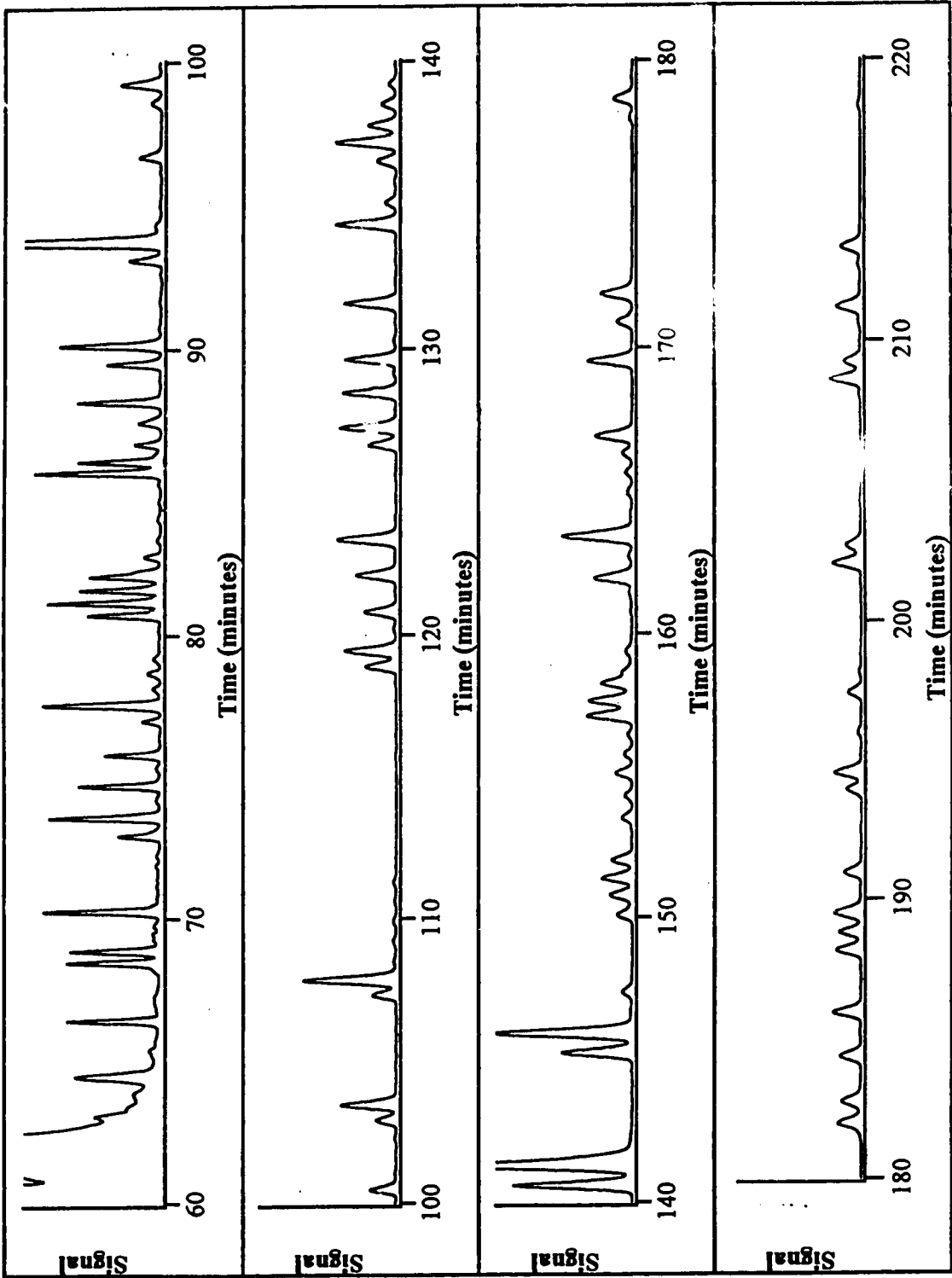
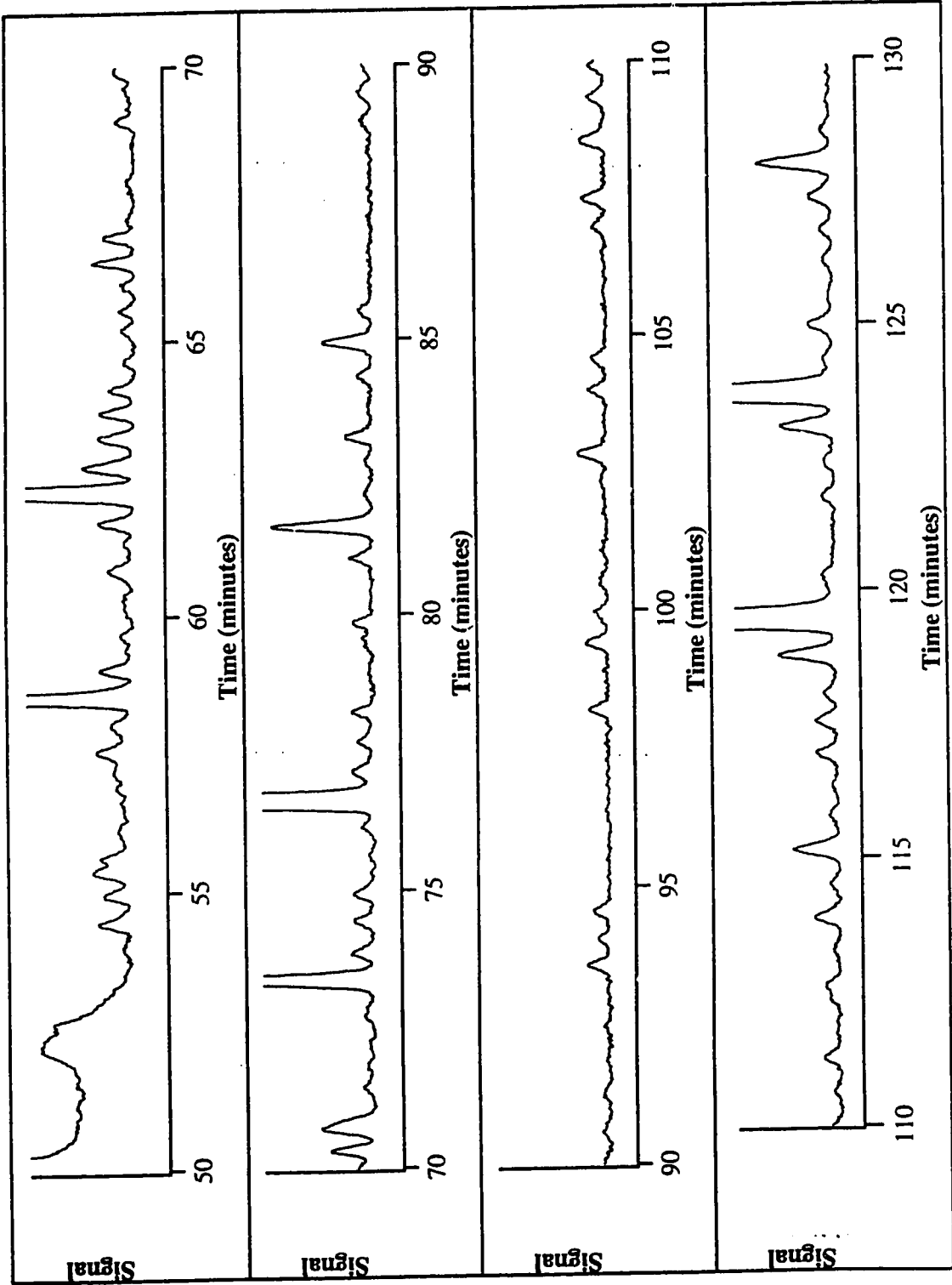


Figure B.3 Template degradation
b. An A terminated sample.

b.



References

1. J.A. Gough, N.E. Murray, *J.Mol. Biol.*, **166** (1983), 1.
2. S. Mizusawa, S. Nishimura, F. Seela, *Nucl. Acids Res.*, **14** (1986), 1319.
3. M.J. Rocheleau, R.J. Grey, D.Y. Chen, H.R. Harke, N.J. Dovichi, *Electrophoresis*, **13** (1992), 484.
4. H. Lu, E. Arriaga, D.Y. Chen, N.J. Dovichi, *J. Chromatogr.*, in press.